IBM SPSS Statistics Base 20

Nota: Prima di utilizzare queste informazioni e il relativo prodotto, leggere le informazioni generali disponibili in Note legali a pag. 309.

Questa versione si applica a IBM® SPSS® Statistics 20 e a tutte le successive versioni e modifiche fino a eventuali disposizioni contrarie indicate in nuove versioni.

Le schermate dei prodotti Adobe sono state ristampate su autorizzazione di Adobe Systems Incorporated.

Le schermate dei prodotti Microsoft sono state ristampate su autorizzazione di Microsoft Corporation.

Prefazione

IBM® SPSS® Statistics è un sistema completo per l'analisi dei dati. Il modulo aggiuntivo opzionale Base include le tecniche di analisi aggiuntive descritte nel presente manuale. Il modulo aggiuntivo Base deve essere usato con il modulo Core SPSS Statistics in cui è completamente integrato.

Informazioni su Business Analytics di IBM

Il software IBM Business Analytics fornisce informazioni complete, coerenti e accurate a cui i responsabili delle decisioni possono affidarsi per ottimizzare le prestazioni dell'azienda. Un ampio portafoglio di applicazioni di business intelligence, analisi predittiva, gestione delle prestazioni e delle strategie finanziarie e analisi offre una panoramica chiara, istantanea e interattiva delle prestazioni attuali e la possibilità di prevedere i risultati futuri. Utilizzato in combinazione con potenti soluzioni di settore, prassi consolidate e servizi professionali, questo software consente alle aziende di tutte le dimensioni di ottimizzare la produttività, automatizzare le decisioni senza problemi e fornire risultati migliori.

Come parte di questo portafoglio, il software IBM SPSS Predictive Analytics consente alle aziende di prevedere gli eventi futuri e di agire tempestivamente in modo da migliorare i risultati delle attività aziendali. Le aziende, gli enti governativi e le università di tutto il mondo si affidano alla tecnologia IBM SPSS perché rappresenta un vantaggio concorrenziale in termini di attrazione, retention e aumento dei clienti, riducendo al tempo stesso le frodi e limitando i rischi. Incorporando il software IBM SPSS nelle attività quotidiane, le aziende diventano imprese in grado di effettuare previsioni e di gestire e automatizzare le decisioni, per raggiungere gli obiettivi aziendali e vantaggi tangibili sulla concorrenza. Per ulteriori informazioni o per contattare un rappresentante, visitare il sito *http://www.ibm.com/spss*.

Supporto tecnico

Ai clienti che richiedono la manutenzione, viene messo a disposizione un servizio di supporto tecnico. I clienti possono contattare il supporto tecnico per richiedere assistenza per l'utilizzo dei prodotti IBM Corp. o per l'installazione di uno degli ambienti hardware supportati. Per contattare il supporto tecnico, visitare il sito Web di IBM Corp. all'indirizzo *http://www.ibm.com/support*. Ricordare che durante la richiesta di assistenza sarà necessario fornire i dati di identificazione personali, i dati relativi alla propria società e il numero del contratto di manutenzione.

Supporto tecnico per studenti

Gli studenti che utilizzano una versione accademica o grad pack di qualsiasi prodotto software IBM SPSS sono pregati di utilizzare le apposite pagine online per studenti Solutions for Education (*http://www.ibm.com/spss/rd/students/*). Gli studenti che utilizzano una copia del software IBM SPSS fornita dall'università, sono pregati di contattare il coordinatore del prodotto IBM SPSS presso l'università.

Servizio clienti

Per eventuali chiarimenti in merito alla spedizione o al proprio conto, rivolgersi alla sede locale. Tenere presente che sarà necessario fornire il numero di serie.

Corsi di formazione

IBM Corp. organizza corsi di formazione pubblici e onsite che includono esercitazioni pratiche. Tali corsi si terranno periodicamente nelle principali città. Per ulteriori informazioni su questi seminari, andare a *http://www.ibm.com/software/analytics/spss/training*.

Pubblicazioni aggiuntive

I documenti *SPSS Statistics: Guide to Data Analysis*, *SPSS Statistics: Statistical Procedures Companion* e *SPSS Statistics: Advanced Statistical Procedures Companion*, scritti da Marija Norušis e pubblicati da Prentice Hall sono disponibili come materiale supplementare consigliato. Queste pubblicazioni descrivono le procedure statistiche nei moduli SPSS Statistics Base, Advanced Statistics e Regression. Utili sia come guida iniziale all'analisi dei dati che per applicazioni avanzate, questi manuali consentono di ottimizzare l'utilizzo delle funzionalità presenti nell'offerta IBM® SPSS® Statistics. Per ulteriori informazioni, inclusi contenuti delle pubblicazioni e capitoli di esempio, visitare il sito Web dell'autrice: *http://www.norusis.com*

Contenuto

Capitolo
1

Informazioni sui dati

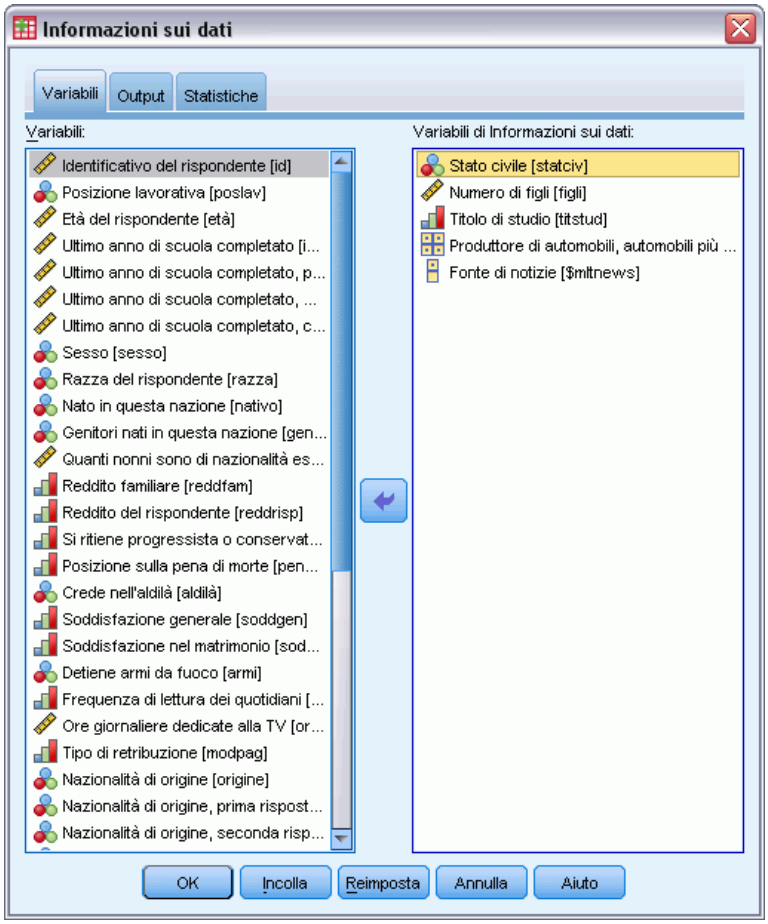
Le informazioni sui dati restituiscono informazioni del dizionario, ad esempio nomi di variabili, etichette di variabili e di valori o valori mancanti, e statistiche riassuntive per alcune o tutte le variabili e gli insiemi a risposta multipla presenti nell'insieme di dati attivo. Per le variabili nominali e ordinali e gli insiemi a risposta multipla, le statistiche riassuntive includono conteggi e percentuali. Per le variabili di scala, le statistiche riassuntive includono media, deviazione standard e quartili.

Nota: Informazioni sui dati ignorano lo stato del file distinto. Sono inclusi i gruppi di file distinti creati per i valori mancanti ad assegnazione multipla (disponibili nel modulo aggiuntivo Missing Values).

Per ottenere le informazioni sui dati

- Dai menu, scegliere:
  Analizza > Report > Informazioni sui dati
- Fare clic sulla scheda Variabili.

Figura 1-1
Finestra di dialogo Informazioni sui dati, scheda Variabili



► Selezionare una o più variabili e/o uno o più insiemi a risposta multipla.

Se lo si desidera, è possibile:

- Controllare le informazioni sulle variabili visualizzate.
- Controllare le statistiche visualizzate (o escludere tutte le statistiche riassuntive).
- Controllare l'ordine in cui vengono visualizzati insiemi a risposta multipla e variabili.
- Modificare il livello di misurazione per le variabili nell'elenco di origine in modo tale da modificare le statistiche riassuntive visualizzate. Per ulteriori informazioni, vedere l'argomento Scheda Informazioni sui dati - Statistiche a pag. 6.

Modifica del livello di misurazione

È possibile modificare temporaneamente il livello di misurazione per le variabili. Non è possibile modificare il livello di misurazione per gli insiemi a risposta multipla, che vengono sempre trattati come nominali)

► Fare clic con il pulsante destro del mouse su una variabile nell'elenco sorgente.

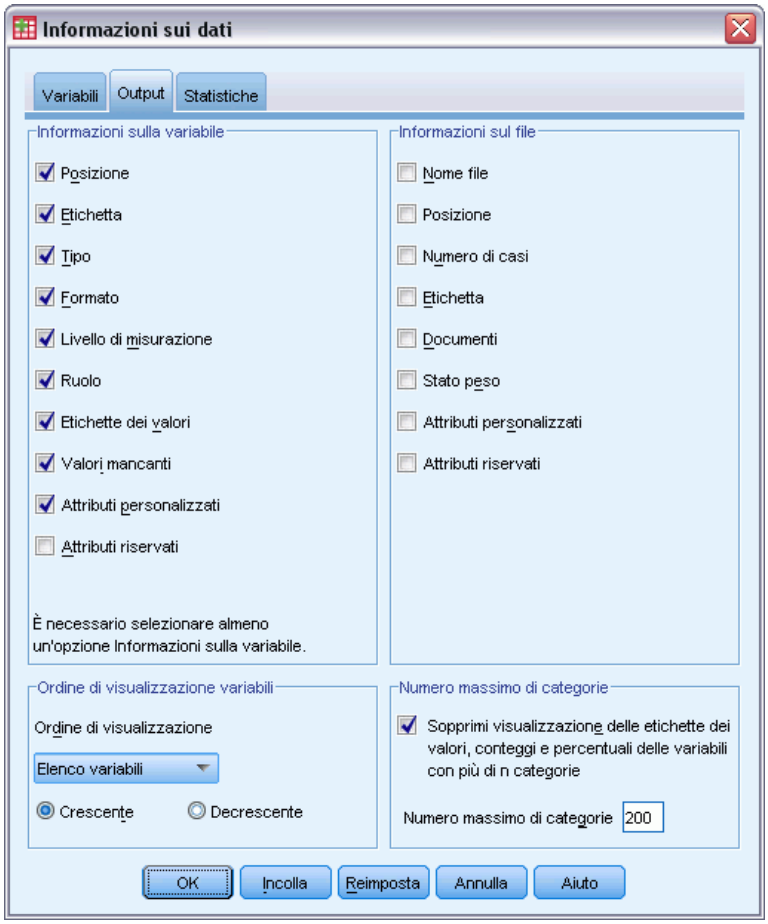
▶ Scegliere un livello di misurazione dal menu di scelta rapida popup.

In questo modo il livello di misurazione viene temporaneamente modificato. Da un punto di vista pratico, è utile solo per le variabili numeriche. Il livello di misurazione per le variabili stringa è limitato a nominale o ordinale, entrambi trattati allo stesso modo dalla procedura Informazioni sui dati.

Scheda Output della finestra Informazioni sui dati

La scheda Output controlla le informazioni sulle variabili incluse per ciascuna variabile e ciascun insieme a risposta multipla, l'ordine in cui variabili e insiemi a risposta multipla vengono visualizzati e il contenuto della tabella delle informazioni facoltative sui file.

Figura 1-2
Finestra di dialogo Informazioni sui dati, scheda Output



Informazioni sulla variabile

Controlla le informazioni del dizionario visualizzate per ciascuna variabile.

Posizione. Intero che rappresenta la posizione della variabile nell'ordine del file. Non è disponibile per gli insiemi a risposta multipla.

Etichetta. Etichetta descrittiva associata alla variabile o all'insieme a risposta multipla.

Tipo. Tipo di dati fondamentale. Può essere *Numerico*, *Stringa* o *Insieme a risposta multipla*.

Formato. Formato di visualizzazione della variabile, ad esempio *A4*, *F8.2* o *DATE11*. Non è disponibile per gli insiemi a risposta multipla.

Livello di misurazione. I valori possibili sono *Nominale*, *Ordinale*, *Scala* e *Sconosciuto*. Il valore visualizzato è il livello di misurazione memorizzato nel dizionario e non è influenzato da alcuna variazione temporanea del livello di misurazione dovuta alla modifica del livello nell'elenco delle variabili sorgente della scheda Variabili. Non è disponibile per gli insiemi a risposta multipla.

Nota: il livello di misurazione per le variabili numeriche può essere "sconosciuto" prima del primo ciclo di dati poiché tale livello non è stato ancora impostato esplicitamente, ad esempio quando i dati vengono letti da una sorgente esterna o le variabili sono appena state create.

Ruolo. Alcune finestre di dialogo supportano la capacità di pre-selezionare le variabili per l'analisi in base a dei ruoli definiti.

Etichette dei valori. Etichette descrittive associate a valori di dati specifici.

- Se nella scheda Statistiche è selezionata l'opzione Conteggio o Percentuale, le etichette dei valori definiti sono incluse nell'output anche se l'opzione Etichette dei valori non è stata selezionata.
- Per gli insiemi a dicotomie multiple, le "etichette dei valori" sono le etichette delle variabili per le variabili elementari presenti nell'insieme o le etichette dei valori conteggiati, in base alla definizione dell'insieme.

Valori mancanti. Valori mancanti definiti dall'utente. Se nella scheda Statistiche è selezionata l'opzione Conteggio o Percentuale, le etichette dei valori definiti sono incluse nell'output anche se l'opzione Valori mancanti non è stata selezionata. Non è disponibile per gli insiemi a risposta multipla.

Attributi personalizzati. Attributi delle variabili definite dall'utente. L'output include sia i nomi sia i valori di tutti gli attributi personalizzati associati a ciascuna variabile. Non è disponibile per gli insiemi a risposta multipla.

Attributi riservati. Attributi riservati delle variabili di sistema. È possibile visualizzare gli attributi di sistema, ma non modificarli. I nomi degli attributi di sistema iniziano con il simbolo del dollaro ($) . Gli attributi non di visualizzazione, che hanno nomi che iniziano con "@" o "$@", non sono inclusi. L'output include sia i nomi sia i valori di tutti gli attributi di sistema associati a ciascuna variabile. Non è disponibile per gli insiemi a risposta multipla.

Informazioni sul file

La tabella delle informazioni opzionali sul file può includere i seguenti attributi:

Nome file. Nome del file di dati di IBM® SPSS® Statistics. Se l'insieme di dati non è mai stato salvato in formato SPSS Statistics, non esiste alcun nome di file di dati. Se non è visualizzato un nome di file nella barra del titolo della finestra Editor dei dati, l'insieme di dati attivo non ha un nome di file.

Posizione. Percorso della directory (cartella) del file di dati di SPSS Statistics. Se l'insieme di dati non è mai stato salvato in formato SPSS Statistics, non esiste alcun percorso.

Numero di casi. Numero di casi presenti nell'insieme di dati attivo. Si tratta del numero totale di casi, compresi gli eventuali casi esclusi dalle statistiche riassuntive a causa delle condizioni di filtro.

Etichetta. Etichetta del file (se presente) definita dal comando `FILE LABEL`.

Documenti. Testo del documento del file di dati.

Stato della ponderazione. Se il peso è attivo, viene visualizzato il nome della variabile di peso.

Attributi personalizzati. Attributi del file di dati definiti dall'utente. Gli attributi del file di dati vengono definiti con il comando `DATAFILE ATTRIBUTE`.

Attributi riservati. Attributi riservati del file di dati di sistema. È possibile visualizzare gli attributi di sistema, ma non modificarli. I nomi degli attributi di sistema iniziano con il simbolo del dollaro ($) . Gli attributi non di visualizzazione, che hanno nomi che iniziano con "@" o "$@", non sono inclusi. L'output include sia i nomi che i valori di tutti gli attributi del file di dati di sistema.

Ordine di visualizzazione variabili

Sono disponibili le seguenti opzioni per il controllo dell'ordine in cui vengono visualizzati gli insiemi a risposta multipla e le variabili.

Alfabetico. Ordine alfabetico per nome di variabile.

File. Ordine in cui le variabili appaiono nell'insieme di dati (l'ordine in cui sono visualizzate nell'Editor dei dati). In ordine crescente, con gli insiemi a risposta multipla visualizzati per ultimi, dopo tutte le variabili selezionate.

Livello di misurazione. Ordina per livello di misurazione. Vengono creati quattro gruppi di ordinamento: nominale, ordinale, scala e sconosciuto. Gli insiemi a risposta multipla vengono trattati come nominali.

Nota: il livello di misurazione per le variabili numeriche può essere "sconosciuto" prima del primo ciclo di dati poiché tale livello non è stato ancora impostato esplicitamente, ad esempio quando i dati vengono letti da una sorgente esterna o le variabili sono appena state create.

Elenco variabili. Ordine in cui le variabili e gli insiemi a risposta multipla appaiono nell'elenco delle variabili selezionate della scheda Variabili.

Nome attributo personalizzato. L'elenco delle opzioni di ordinamento include anche i nomi degli attributi personalizzati delle variabili definite dall'utente. In ordine crescente, con le variabili senza attributi per prime, seguite dalle variabili che hanno un attributo ma senza valori definiti per lo stesso, seguite dalle variabili con valori definiti per l'attributo e valori in ordine alfabetico.

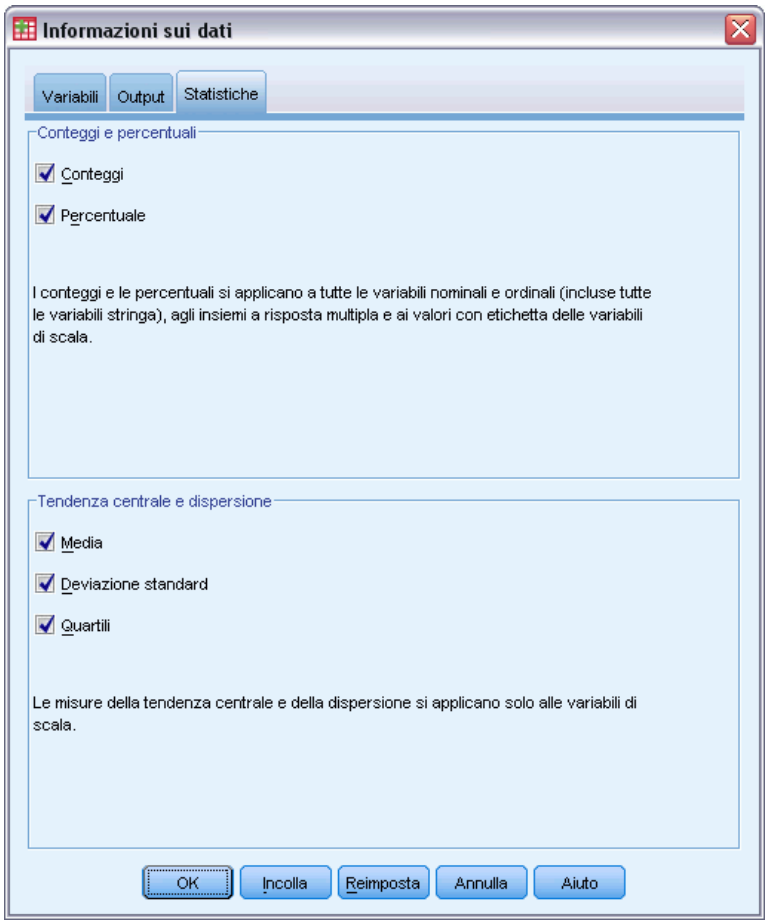
Numero massimo di categorie

Se l'output include etichette di valori, conteggi o percentuali per ogni valore univoco, è possibile eliminare questa informazione dalla tabella se il numero di valori supera il valore specificato. Per impostazione predefinita, questa informazione non viene inserita se il numero di valori univoci per la variabile supera 200.

Scheda Informazioni sui dati - Statistiche

La scheda Statistiche consente di controllare le statistiche riassuntive incluse nell'output o di eliminare completamente la visualizzazione delle statistiche riassuntive.

Figura 1-3
Finestra di dialogo Informazioni sui dati, scheda Statistiche



Conteggi e percentuali

Per le variabili nominali e ordinali, gli insiemi a risposta multipla e i valori con etichetta delle variabili di scala, le statistiche riassuntive sono:

Conteggio. Il conteggio o numero di casi con ogni valore (o intervallo di valori) di una variabile.

Percentuale. La percentuale di casi con un valore particolare.

Tendenza centrale e dispersione

Per le variabili di scala, le statistiche disponibili sono:

Media. Una misura di tendenza centrale. La somma dei valori di tutte le osservazioni divisa per il numero di osservazioni. Viene anche detta media aritmetica.

Deviazione standard. La radice quadrata della varianza. La deviazione standard è una misura della dispersione intorno alla media espressa nella stessa unità di misura delle osservazioni. In una distribuzione normale, il 68% dei casi rientra in una deviazione standard della media e il 95% dei casi rientra in due deviazioni standard. Se, ad esempio, in una popolazione con distribuzione normale l'età media fosse 45 e la deviazione standard 10, il 95% dei casi cadrebbe fra 25 e 65 anni.

Quartili. Mostra i valori corrispondenti al 25°, 50° e 75° percentile.

Nota: è possibile modificare temporaneamente il livello di misurazione associato a una variabile (e quindi modificare le statistiche riassuntive visualizzate per quella variabile) nell'elenco variabili sorgente della scheda Variabili.

Capitolo
2

Frequenze

La procedura Frequenze consente di ottenere statistiche e rappresentazioni grafiche che risultano utili per la descrizione di molti tipi di variabili. La procedura Frequenza offre un'ottima opportunità per iniziare ad osservare i dati.

Per ottenere un rapporto e un grafico a barre delle frequenze è possibile disporre i singoli valori in ordine crescente o decrescente oppure ordinare le categorie in base alle rispettive frequenze. Il rapporto sulle frequenze può essere eliminato se una variabile ha molti valori distinti. È possibile etichettare i grafici con frequenze (default) o percentuali.

Esempio. Qual è la distribuzione dei clienti di un'azienda per tipo di industria? Dall'output, si nota che il 37,5% dei clienti fa parte di enti governativi, il 24,9% fa parte di società, il 28,1% di istituzioni accademiche e il 9,4% del settore sanitario. Per i dati quantitativi e continui, ad esempio il fatturato, si può notare che la vendita media del prodotto è pari a €. 3.576 con una deviazione standard di €. 1.078.

Statistiche e grafici. Conteggi di frequenza, percentuali, percentuali cumulate, media, mediana, moda, somma, deviazione standard, varianza, intervallo, valori minimo e massimo, errore standard della media, asimmetria e curtosi (entrambe con errori standard), quartili, percentili definiti dall'utente, grafici a barre, grafici a torta e istogrammi.

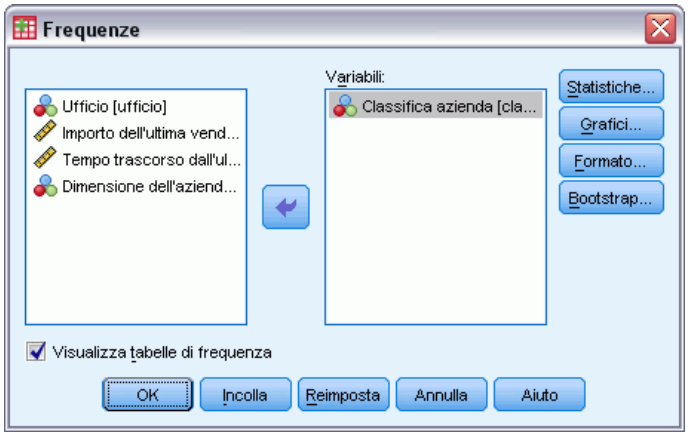
Dati. Utilizzare codici numerici o stringhe per codificare le variabili categoriali (misure di livello nominale o ordinale).

Assunzioni. I riepiloghi e le percentuali forniscono un'utile descrizione dei dati provenienti da qualsiasi distribuzione, in particolare per le variabili con categorie ordinate o non ordinate. La maggior parte delle statistiche riassuntive, ad esempio la media e la deviazione standard, si basano sulla normale teoria e sono idonee per variabili quantitative con distribuzioni simmetriche. Le statistiche robuste, ad esempio la media, i quartili e i percentili, sono idonee per variabili quantitative rispondenti o meno all'ipotesi di normalità.

Per ottenere le tabelle di frequenza

- Dai menu, scegliere:
  Analizza > Statistiche descrittive > Frequenze...

Figura 2-1
Finestra di dialogo principale Frequenze



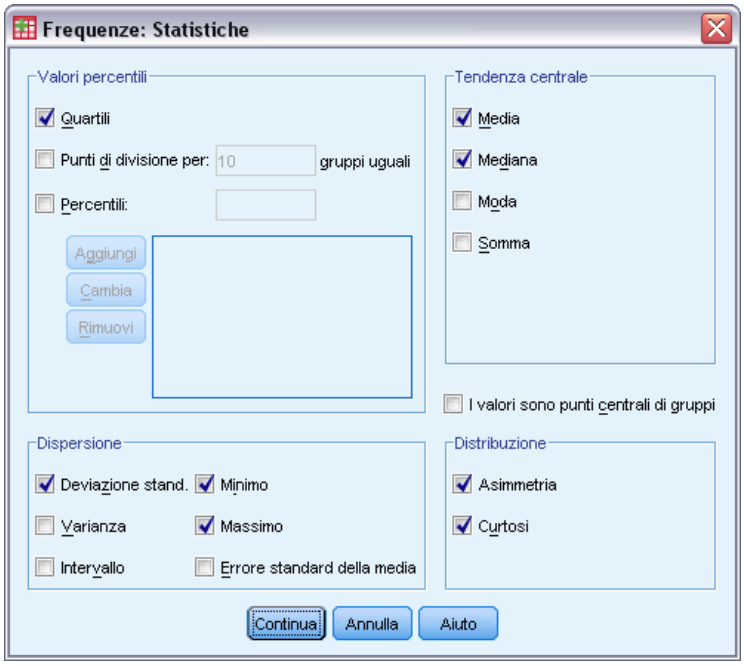
- Selezionare una o più variabili categoriali o quantitative.

Se lo si desidera, è possibile:

- Fare clic su Statistiche per ottenere statistiche descrittive per le variabili quantitative.
- Fare clic su Grafici per ottenere grafici a barre, grafici a torta e istogrammi.
- Fare clic su Formato per stabilire l'ordine in cui visualizzare i risultati.

Frequenze: Statistiche

Figura 2-2
Finestra di dialogo Frequenze: Statistiche



Valori percentili. Valori di una variabile quantitativa che suddividono i dati ordinati in due gruppi in modo da visualizzare una percentuale sopra e una sotto. I quartili (il 25°, 50° e 75° percentile) suddividono le osservazioni in quattro gruppi di dimensioni uguali. Se si desidera ottenere un numero di gruppi uguali diverso da quattro, selezionare Punti di divisione per gruppi uguali. È inoltre possibile specificare i singoli percentili, ad esempio il 95° percentile, ovvero il valore al di sotto del quale ricade il 95% delle osservazioni.

Tendenza centrale. Le statistiche che descrivono la posizione della distribuzione includono media, mediana, moda e somma di tutti i valori.

- **Media.** Una misura di tendenza centrale. La somma dei valori di tutte le osservazioni divisa per il numero di osservazioni. Viene anche detta media aritmetica.
- **Mediana.** È il valore sopra il quale e sotto il quale ricade la metà dei casi, il 50-esimo percentile. Se il numero di casi è pari, la mediana è pari alla media dei due casi centrali quando questi sono ordinati secondo l'ordine ascendente o discendente. La mediana è una misura di tendenza centrale non sensibile ai valori anomali, a differenza della media che può essere influenzata da valori eccezionalmente bassi o alti.
- **Modalità.** Il valore che ricorre più frequentemente. Se più valori condividono la maggiore ricorrenza, ognuno di essi è una moda. La procedura Frequenze riporta solo la più piccola delle mode.
- **Somma.** La somma o il totale di tutti i valori non mancanti di tutti i casi.

Dispersione. Le statistiche che misurano l'entità della variazione o della variabilità dei dati includono deviazione standard, varianza, intervallo, valore minimo e massimo ed errore standard della media.

- **Deviazione stand..** La radice quadrata della varianza. La deviazione standard è una misura della dispersione intorno alla media espressa nella stessa unità di misura delle osservazioni. In una distribuzione normale, il 68% dei casi rientra in una deviazione standard della media e il 95% dei casi rientra in due deviazioni standard. Se, ad esempio, in una popolazione con distribuzione normale l'età media fosse 45 e la deviazione standard 10, il 95% dei casi cadrebbe fra 25 e 65 anni.
- **Varianza.** Una misura della dispersione dei valori intorno alla media. È calcolata come somma dei quadrati degli scostamenti dalla media, divisa per il numero totale delle osservazioni valide meno 1. La varianza è espressa in quadrati dell'unità di misura della variabile.
- **Intervallo.** La differenza tra il valore massimo ed il valore minimo di una variabile numerica.
- **Minimo.** Il valore più basso assunto da una variabile numerica.
- **Massimo.** Il valore più alto di una variabile numerica.
- **E. S. media.** Una misura di quanto può variare il valore della media da campione a campione per campioni estratti dalla stessa distribuzione. Può essere utilizzata per confrontare genericamente la media osservata rispetto a un valore ipotizzato (ovvero, è possibile concludere che i due valori sono diversi se il rapporto della differenza rispetto all'errore standard è inferiore a -2 o maggiore di +2).

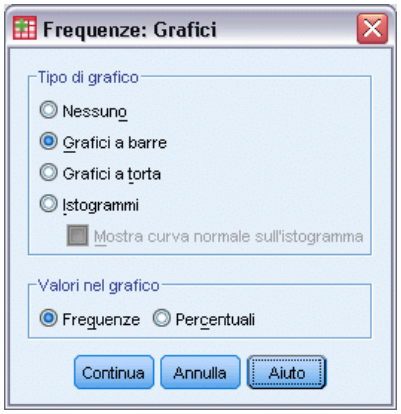
Distribuzione. L'asimmetria e la curtosi sono statistiche che descrivono la forma e la simmetria della distribuzione. Queste statistiche vengono visualizzate con i relativi errori standard.

- **Asimmetria.** Una misura dell'asimmetria di una distribuzione. La distribuzione normale è simmetrica e ha un valore di asimmetria pari a 0. Una distribuzione con una notevole asimmetria positiva ha una lunga coda a destra. Una distribuzione con asimmetria negativa ha una coda a sinistra. In generale un'asimmetria con valore più che doppio dell'errore standard indica lo scostamento dalla normale simmetria.
- **Curtosi.** Una misura di quanto le osservazioni si trovino raggruppate nelle code. Per la distribuzione normale, il valore della statistica di curtosi è zero. Una curtosi positiva indica che, rispetto a una distribuzione normale, le osservazioni sono più raggruppate intorno al centro della distribuzione e hanno code più sottili fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione leptocurtica sono più spesse rispetto a una distribuzione normale. Una curtosi negativa indica che, rispetto a una distribuzione normale, le osservazioni sono meno raggruppate e hanno code più spesse fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione platicurtica sono più sottili rispetto a una distribuzione normale.

I valori sono punti centrali di gruppi. Se i valori dei dati sono punti centrali di gruppi (ad esempio, l'età delle persone sulla trentina è codificata come 35), selezionare questa opzione per valutare la media e i percentili per i dati originali non raggruppati.

Frequenze: Grafici

Figura 2-3
Finestra di dialogo Frequenze: Grafici.



Tipo di grafico. I grafici a torta mostrano il contributo delle parti all'intero grafico. Ogni sezione di un grafico a torta corrisponde a un gruppo definito da una singola variabile di raggruppamento. Nei grafici a barre il conteggio relativo a ciascun valore o categoria viene rappresentato come una barra distinta, in modo da poter confrontare visivamente le categorie. Anche gli istogrammi contengono barre, che però sono tracciate lungo una scala per intervalli uguali. L'altezza di ogni barra rappresenta il conteggio dei valori di una variabile quantitativa che rientra nell'intervallo. Nell'istogramma vengono indicati la forma, il centro e la variabilità della distribuzione. Una curva normale sovrapposta all'istogramma consente di valutare se i dati sono distribuiti normalmente.

Valori nel grafico. Per i grafici a barre, l'asse di scala può essere etichettato in base ai conteggi o alle percentuali di frequenza.

Frequenze: Formato

Figura 2-4
Finestra di dialogo Frequenze: Formato

Ordina per. La tabella di frequenza può essere disposta in base ai valori effettivi dei dati oppure in base al conteggio (frequenza di ricorrenza) di tali valori, in ordine crescente o decrescente. Se, tuttavia, si desidera ottenere un istogramma o i percentili, si presume che la variabile sia quantitativa e i suoi valori vengano visualizzati in ordine crescente.

Variabili multiple. Se si producono tabelle di statistiche per variabili multiple, è possibile visualizzare tutte le variabili in un'unica tabella (Confronta variabili) o visualizzare una tabella distinta di statistiche per ciascuna variabile (Output per variabili).

Sopprimi le tabelle con più di n modalità. Questa opzione consente di disattivare la visualizzazione delle tabelle che includono un numero di valori maggiore di quello specificato.

Capitolo
3

Descrittive

La procedura Descrittive consente di visualizzare statistiche riassuntive univariate per diverse variabili incluse nella stessa tabella e di calcolare i valori standardizzati (punteggi *z*). È possibile ordinare le variabili in base alle dimensioni delle rispettive medie (in ordine crescente o decrescente), in ordine alfabetico oppure nell'ordine in cui sono state selezionate (impostazione predefinita).

I punteggi *z* salvati vengono aggiunti ai dati nell'Editor dei dati e sono disponibili per la creazione di grafici, elenchi di dati e analisi. Quando le variabili vengono registrate in unità diverse (ad esempio, prodotto interno lordo pro capite e percentuale di alfabetizzazione), una trasformazione dei punteggi *z* consente di posizionare le variabili su una scala comune per facilitarne il confronto visivo.

Esempio. Se ciascun caso incluso nei dati contiene i totali delle vendite giornaliere relativi a ciascun agente di vendita (ad esempio, una voce per Roberto, una per Carlo e una per Bruno), registrati ogni giorno per diversi mesi, la procedura Descrittive consente di calcolare la media delle vendite giornaliere per ogni agente e di ordinare i risultati dalla media di vendita maggiore alla minore.

Statistiche. Dimensioni del campione, media, valore minimo e massimo, deviazione standard, varianza, intervallo, somma, errore standard della media, curtosi e asimmetria degli errori standard.

Dati. Utilizzare variabili numeriche dopo averle valutate graficamente per registrare errori, valori anomali e anomalie distributive. La procedura Descrittive risulta molto utile quando si utilizzano file di grandi dimensioni (migliaia di casi).

Assunzioni. La maggior parte delle statistiche disponibili (compresi i punteggi *z*) si fondano sulla teoria di normalità e possono essere utilizzate per le variabili quantitative (misurazioni a livello di intervallo o di rapporto) con distribuzioni simmetriche. Evitare variabili con categorie non ordinate o distribuzioni asimmetriche. La distribuzione dei punteggi *z* ha la stessa forma di quella dei dati originali. Pertanto, il calcolo dei punteggi *z* non rappresenta una soluzione per dati problematici.

Per ottenere statistiche descrittive

- Dai menu, scegliere:
  Analizza > Statistiche descrittive > Descrittive...

Figura 3-1
Descrittive

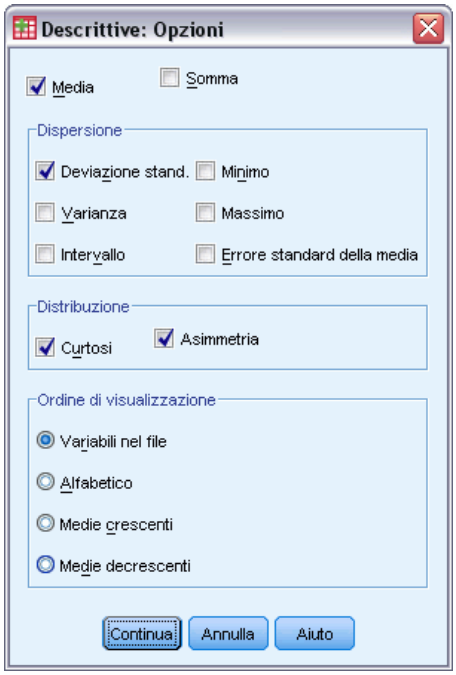
- Selezionare una o più variabili.

Se lo si desidera, è possibile:

- Selezionare Salva valori standardizzati come variabili per salvare i punteggi *z* come nuove variabili.
- Fare clic su Opzioni per ottenere le statistiche e l'ordine di visualizzazione facoltativi.

Descrittive: Opzioni

Figura 3-2
Finestra di dialogo Descrittive: Opzioni



Media e somma. Per impostazione predefinita, viene visualizzata la media o la media aritmetica.

Dispersione. Le statistiche che misurano la dispersione o la variazione dei dati includono deviazione standard, varianza, intervallo, valore minimo e massimo ed errore standard della media.

- **Deviazione stand..** La radice quadrata della varianza. La deviazione standard è una misura della dispersione intorno alla media espressa nella stessa unità di misura delle osservazioni. In una distribuzione normale, il 68% dei casi rientra in una deviazione standard della media e il 95% dei casi rientra in due deviazioni standard. Se, ad esempio, in una popolazione con distribuzione normale l'età media fosse 45 e la deviazione standard 10, il 95% dei casi cadrebbe fra 25 e 65 anni.
- **Varianza.** Una misura della dispersione dei valori intorno alla media. È calcolata come somma dei quadrati degli scostamenti dalla media, divisa per il numero totale delle osservazioni valide meno 1. La varianza è espressa in quadrati dell'unità di misura della variabile.
- **Intervallo.** La differenza tra il valore massimo ed il valore minimo di una variabile numerica.
- **Minimo.** Il valore più basso assunto da una variabile numerica.
- **Massimo.** Il valore più alto di una variabile numerica.
- **Errore standard della media.** Una misura di quanto può variare il valore della media da campione a campione per campioni estratti dalla stessa distribuzione. Può essere utilizzata per confrontare genericamente la media osservata rispetto a un valore ipotizzato (ovvero, è possibile concludere che i due valori sono diversi se il rapporto della differenza rispetto all'errore standard è inferiore a -2 o maggiore di +2).

Distribuzione. Curtosi e asimmetria sono statistiche che caratterizzano la forma e la simmetria della distribuzione. Queste statistiche vengono visualizzate con i relativi errori standard.

- **Curtosi.** Una misura di quanto le osservazioni si trovino raggruppate nelle code. Per la distribuzione normale, il valore della statistica di curtosi è zero. Una curtosi positiva indica che, rispetto a una distribuzione normale, le osservazioni sono più raggruppate intorno al centro della distribuzione e hanno code più sottili fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione leptocurtica sono più spesse rispetto a una distribuzione normale. Una curtosi negativa indica che, rispetto a una distribuzione normale, le osservazioni sono meno raggruppate e hanno code più spesse fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione platicurtica sono più sottili rispetto a una distribuzione normale.
- **Asimmetria.** Una misura dell'asimmetria di una distribuzione. La distribuzione normale è simmetrica e ha un valore di asimmetria pari a 0. Una distribuzione con una notevole asimmetria positiva ha una lunga coda a destra. Una distribuzione con asimmetria negativa ha una coda a sinistra. In generale un'asimmetria con valore più che doppio dell'errore standard indica lo scostamento dalla normale simmetria.

Ordine di visualizzazione. Per impostazione predefinita, le variabili vengono visualizzate nell’ordine in cui vengono selezionate. È inoltre possibile visualizzare le variabili in ordine alfabetico, per media crescente o per media decrescente.

Funzioni aggiuntive del comando DESCRIPTIVES

Il linguaggio della sintassi dei comandi consente inoltre di:

- Salvare i punteggi standardizzati (punteggi *z*) per alcune ma non per tutte le variabili (con il sottocomando `VARIABLES`).
- Specificare i nomi delle nuove variabili che contengono i punteggi standardizzati (con il sottocomando `VARIABLES`).
- Escludere dall'analisi i casi con valori mancanti per qualsiasi variabile (con il sottocomando `MISSING`).
- Ordinare le variabili visualizzate in base al valore di una statistica, non solo in base alla media (con il sottocomando `SORT`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
4

Esplora

La procedura Esplora produce 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 alternativa è possibile utilizzare test non parametrici.

Esempio. Si consideri la distribuzione dei tempi in cui quattro gruppi di ratti imparano a uscire da un labirinto. Per ciascuno dei quattro gruppi, è possibile verificare se la distribuzione dei tempi è approssimativamente normale e se i quattro valori di varianza sono uguali. È inoltre possibile identificare i casi con i cinque tempi più lunghi e i cinque tempi più brevi. I grafici a scatole e i grafici ramo-foglia riassumono graficamente la distribuzione dei tempi di apprendimento per ciascun gruppo.

Statistiche e grafici. Media, mediana, media 5% trim, errore standard, varianza, deviazione standard, valore minimo e massimo, intervallo, distanza interquartilica, asimmetria e curtosi e i relativi errori standard, intervallo di confidenza per la media (e il livello di confidenza specificato), percentili, stimatore M di Huber, stimatore M di Andrew, stimatore M decrescente di Hampel, stimatore di Tukey a doppio peso, i cinque valori maggiori e i cinque valori minori, il test di Kolmogorov-Smirnov con il livello di significatività di Lilliefors per il test della normalità e il test di Shapiro-Wilk. Grafici a scatole, grafici ramo-foglia, istogrammi, grafici di normalità e grafici di variabilità contro intensità con test di Levene e trasformazioni.

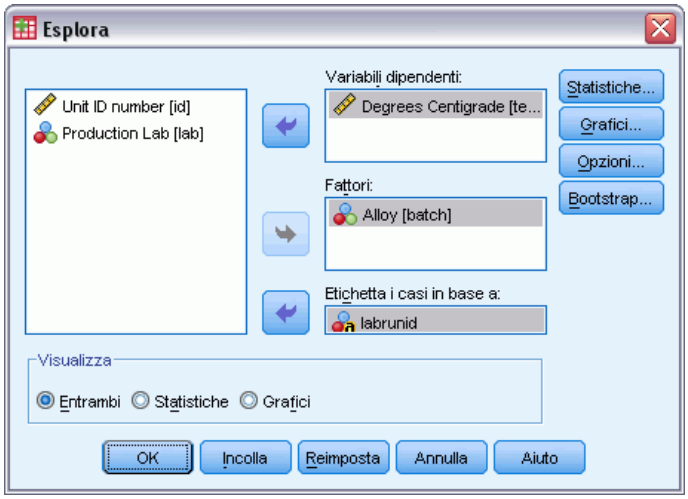
Dati. La procedura Esplora può essere utilizzata per le variabili quantitative (livello di misurazione per intervallo o per rapporto). La variabile fattore, utilizzata per suddividere i dati in gruppi di casi, deve includere un numero ragionevole di valori distinti (categorie). Tali valori possono essere stringhe corte o numerici. La variabile etichetta di caso, utilizzata per etichettare i valori anomali in grafici a scatole, può essere una variabile stringa corta, stringa lunga (i primi 15 byte) o numerica.

Assunzioni. La distribuzione dei dati non deve essere necessariamente simmetrica o normale.

Per esplorare i dati

- Dai menu, scegliere:
  Analizza > Statistiche descrittive > Esplora...

Figura 4-1
Finestra di dialogo Esplora



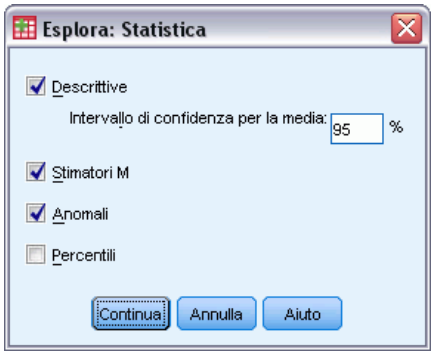
▶ Selezionare una o più variabili dipendenti.

Se lo si desidera, è possibile:

- Selezionare una o più variabili fattore i cui valori definiranno i gruppi di casi.
- Selezionare una variabile di identificazione per etichettare i casi.
- Fare clic su Statistiche per ottenere stimatori robusti, valori anomali, percentili e tabelle di frequenza.
- Fare clic su Grafici per ottenere istogrammi, grafici e test di probabilità normale e grafici di variabilità contro intensità con test di Levene.
- Fare clic su Opzioni per ottenere il trattamento dei valori mancanti.

Esplora: Statistica

Figura 4-2
Finestra di dialogo Esplora: Statistica



Descrittive. Queste misure di tendenza centrale e di dispersione vengono visualizzate per impostazione predefinita. Le misure di tendenza centrale indicano la posizione della distribuzione e includono la media, la mediana e la media 5% trim. Le misure di dispersione mostrano la

dissimilarità dei valori e includono errore standard, varianza, deviazione standard, valore minimo e massimo, intervallo e distanza interquartilica. Le statistiche descrittive includono anche le misure della forma della distribuzione; l'asimmetria e la curtosi vengono visualizzate con i rispettivi errori standard. Viene visualizzato anche l'intervallo di confidenza al 95% per la media. È possibile specificare un diverso livello di confidenza.

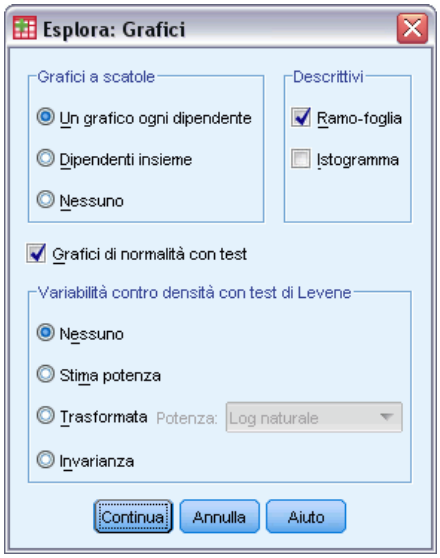
Stimatori M. Alternative valide alla media e alla mediana del campione per la valutazione della posizione. Gli stimatori calcolati differiscono per il peso applicato ai casi. Verranno visualizzati lo stimatore M di Huber, lo stimatore M di Andrews, lo stimatore M decrescente di Hampel e lo stimatore di Tukey a doppio peso.

Valori anomali. Consente di visualizzare i cinque valori maggiori e i cinque valori minori con le etichette dei casi.

Percentili. Consente di visualizzare i valori del 5°, 10°, 25°,50°, 75°, 90° e 95° percentile.

Esplora: Grafici

Figura 4-3
Finestra di dialogo Esplora: Grafici



Grafici a scatole. Queste alternative controllano la visualizzazione dei grafici a scatole quando sono presenti più variabili dipendenti. Un grafico ogni dipendente consente di generare una visualizzazione distinta per ciascuna variabile dipendente. All'interno della visualizzazione, vengono visualizzati grafici a scatole per ciascun gruppo definito da una variabile fattore. Dipendenti insieme consente di generare una visualizzazione distinta per ciascun gruppo definito da una variabile fattore. All'interno della visualizzazione compaiono grafici a scatole affiancati per ciascuna variabile dipendente. Questo tipo di grafico risulta particolarmente utile quando le singole variabili rappresentano una caratteristica misurata in tempi diversi.

Descrittive. Nel gruppo Descrittive è possibile scegliere grafici ramo-foglia e istogrammi.

Grafici di normalità con test. Consente di visualizzare grafici di probabilità normale e grafici di probabilità normale detrendizzati. Viene visualizzato il test di Kolmogorov-Smirnov con un livello di significatività di Lilliefors per il test della normalità. Se i pesi non interi sono specificati, la statistica di Shapiro-Wilk viene calcolata quando la dimensione campione pesata è compresa tra 3 e 50. Per pesi interi o non pesi, la statistica viene calcolata quando la dimensione del campione pesato è compresa tra 3 e 5.000.

Variabilità vs. intensità con test di Levene. Consente di controllare la trasformazione dei dati per i grafici di variabilità contro intensità. Per tutti i grafici di variabilità contro intensità vengono visualizzati l'inclinazione della curva di regressione e i test di Levene per l'omogeneità della varianza. Se si seleziona una trasformazione, i test di Levene si baseranno sui dati trasformati. Se non viene selezionata una variabile fattore, non verranno creati grafici di variabilità contro intensità. Stima potenza traccia i logaritmi naturali delle distanze interquartiliche verso i logaritmi naturali delle mediane di tutte le celle e inoltre una stima della potenza necessaria per trasformare i dati in modo da raggiungere varianze uguali in tutte le celle. Un grafico variabilità contro intensità consente di identificare la potenza di una trasformazione per stabilizzare (rendere maggiormente uguale) le varianze nei vari gruppi. Trasformata consente di selezionare un valore di potenza alternativo, seguendo o meno le indicazioni della stima di potenza, e di produrre i grafici dei dati trasformati. La distanza interquartilica e la media dei dati trasformati verranno tracciate in un grafico. Invarianza consente di ottenere grafici relativi ai dati semplici. Equivale a una trasformazione con potenza 1.

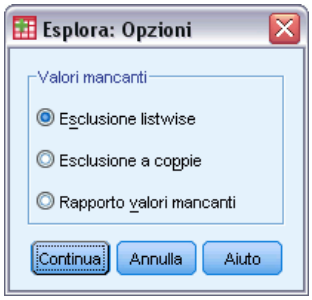
Esplora: potenza necessaria per la trasformazione dei dati

Si tratta delle trasformazioni di potenza per i grafici di variabilità contro intensità. Per trasformare i dati è necessario selezionare la potenza corrispondente. È possibile scegliere una delle seguenti opzioni:

- **Log naturale.** Trasformazione logaritmica naturale. È l'impostazione predefinita.
- **1/radice quadrata.** Per ciascun valore viene calcolato il reciproco della radice quadrata.
- **Reciproco.** Viene calcolato il reciproco di ciascun valore.
- **Radice quadrata.** Viene calcolata la radice quadrata di ciascun valore.
- **Quadrato.** Ciascun valore viene elevato al quadrato.
- **Cubo.** Ciascun valore viene elevato al cubo.

Esplora: Opzioni

Figura 4-4
Finestra di dialogo Esplora: Opzioni



Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile dipendente o fattore verranno esclusi da tutte le analisi. È l'impostazione predefinita.
- **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 supplementare. Le tabelle di frequenza includono categorie per i valori mancanti. I valori mancanti per una variabile fattore vengono inclusi, ma etichettati come mancanti.

Funzioni aggiuntive del comando EXAMINE

La procedura Esplora usa la sintassi del comando `EXAMINE`. Il linguaggio della sintassi dei comandi consente inoltre di:

- Richiedere l'output totale e i grafici oltre all'output e ai grafici per i gruppi definiti dalle variabili di fattore (con il sottocomando `TOTAL`).
- Specificare una scala comune per un gruppo di grafici a scatole (con il sottocomando `SCALE`).
- Specificare le interazioni delle variabili di fattore (con il sottocomando `VARIABLES`).
- Specificare percentili diversi da quelli predefiniti (con il sottocomando `PERCENTILES`).
- Calcolare i percentili utilizzando uno dei cinque metodi (con il sottocomando `PERCENTILES`).
- Specificare una trasformazione di potenza per i grafici di variabilità vs. intensità (con il sottocomando `PLOT`).
- Specificare il numero di valori estremi da visualizzare (con il sottocomando `STATISTICS`).
- Specificare i parametri per i predittori M e i predittori robusti di posizione (con il sottocomando `MESTIMATORS`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
5

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). Ad esempio, se la variabile *sesso* è uno strato per la tabella della variabile *coniugato* (sì, no) rispetto alla variabile *tipo di vita* (ottima, soddisfacente, non soddisfacente), i risultati per la tabella bivariata per le donne vengono elaborati separatamente da quelli per gli uomini e quindi stampati come riquadri in successione.

Esempio. È possibile che i clienti rappresentati da piccole società siano più remunerativi per la vendita di servizi (per esempio addestramenti e consulenze) rispetto ai clienti rappresentati da società di grandi dimensioni? Mediante una tavola di contingenza è possibile scoprire che la maggior parte delle società di piccole dimensioni (con un numero di dipendenti inferiore a 500) fruttano alti profitti per i servizi, mentre la maggior parte delle grandi società (con oltre 2,500 dipendenti) fruttano profitti di scarsa entità.

Statistiche e misure di associazione. Chi-quadrato di Pearson, chi-quadrato del rapporto di verosimiglianza, test di associazione lineare-lineare, test esatto di Fisher, chi-quadrato corretto di Yates, *R* di Pearson, rho di Spearman, coefficiente di contingenza, phi, *V* di Cramér, lambda simmetrica e asimmetrica, tau di Goodman e Kruskal, coefficiente di incertezza, gamma, *D* di Somers, tau-*b* di Kendall, tau-*c* di Kendall, coefficiente eta, kappa di Cohen, stima del rischio relativo, rapporto odd, test di McNemar, statistiche di Cochran e Mantel-Haenszel e statistiche delle proporzioni di colonna.

Dati. Per definire le categorie di ciascuna variabile della tabella, utilizzare i valori di una variabile numerica o stringa (con una lunghezza massima di otto byte). Ad esempio, per la variabile *sesso*, è possibile codificare i dati come 1 e 2 oppure come *maschio* e *femmina*.

Assunzioni. Alcune statistiche e misure assumono categorie ordinate (dati ordinali) o valori quantitativi (dati misurati per intervallo o per rapporto), come indicato nella sezione sulle statistiche. Se le variabili della tabella prevedono categorie non ordinate (dati nominali), sono disponibili altri valori validi. Per le statistiche basate sul chi-quadrato (phi, *V* di Cramér e coefficiente di contingenza), i dati devono essere rappresentati da un campione casuale proveniente da una distribuzione multinomiale.

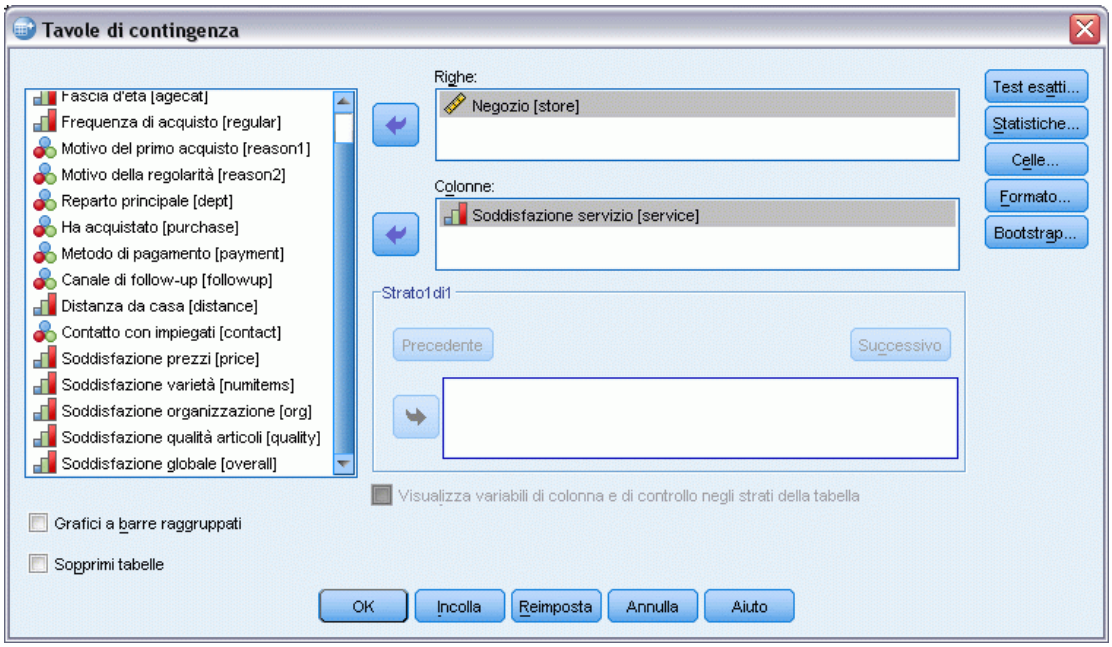
Nota: le variabili ordinali possono essere codici numerici che rappresentano categorie (ad esempio 1 = *basso*, 2 = *medio*, 3 = *alto*) oppure valori di stringa. Si suppone tuttavia che l'ordine alfabetico dei valori di stringa rifletta l'esatto ordine delle categorie. Ad esempio, per una variabile stringa con i valori *basso*, *medio*, *alto*, l'ordine delle categorie viene interpretato come *alto*, *basso*, *medio*,

ma questo non è l'ordine corretto. In generale, per rappresentare i dati ordinali, è più sicuro utilizzare i codici numerici.

Per ottenere tavole di contingenza

▶ Dai menu, scegliere:
Analizza > Statistiche descrittive > Tavole di contingenza...

Figura 5-1
Finestra di dialogo Tavole di contingenza



▶ Selezionare una o più variabili di riga e una o più variabili di colonna.

Se lo si desidera, è possibile:

- Selezionare una o più variabili di controllo.
- Fare clic su Statistiche per ottenere test e misure di associazione per tabelle o sottotabelle bivariate.
- Fare clic suCelle per ottenere valori, percentuali e residui osservati e attesi.
- Fare clic suFormato per controllare l'ordine delle categorie.

Strati nelle tavole di contingenza

Se vengono selezionate una o più variabili di strato, verrà prodotta una tavola di contingenza distinta per ciascuna categoria di ciascuna variabile di strato (variabile di controllo). Ad esempio, se si dispone di una variabile di riga, una variabile di colonna e una variabile di strato con due categorie, si otterrà una tabella bivariata per ciascuna categoria della variabile di strato. Per creare un altro strato di variabili di controllo, fare clic su Successivo. Verranno create sottotabelle per ogni combinazione delle categorie di ciascuna variabile del primo strato con ciascuna variabile del

secondo e così via. Se sono richieste statistiche e misure di associazione, verranno applicate solo alle sottotabelle bivariate.

Tavole di contingenza: grafici a barre raggruppati

Grafici a barre raggruppati. Nei grafici a barre raggruppati è possibile riepilogare i dati relativi a gruppi di casi. È disponibile un gruppo di barre per ciascun valore della variabile specificata in Righe. La variabile che definisce le barre contenute in ogni gruppo è quella specificata in Colonne. Per ciascun valore della variabile è disponibile una serie di barre con colori e motivi diversi. Se in Colonne o Righe si specificano più variabili, verrà prodotto un grafico a barre raggruppato per ciascuna combinazione delle due variabili.

Tavole di contingenza con variabili di strato negli strati della tabella

Visualizza variabili di strato negli strati della tabella. È possibile scegliere di visualizzare le variabili di strato (variabili di controllo) come strati della tabella nella tavola di contingenza. Ciò consente di creare visualizzazioni che mostrano le statistiche globali per le variabili di riga e di colonna e permettono di visualizzare i dettagli delle categorie delle variabili di strato.

L'esempio riportato di seguito utilizza il file di dati *demo.sav* () ed è stato ottenuto come segue:

- Selezionare *Categoria Reddito in migliaia (catredd)* come variabile di riga, *Possiede PDA (pda)* come variabile di colonna e *Livello di istruzione (istruz)* come variabile di strato.
- Selezionare Visualizza variabili di strato negli strati della tabella.
- Selezionare Colonna nella finestra di dialogo secondaria Visualizzazione cella.
- Eseguire la procedura Tavole di contingenza, fare doppio clic sulla tavola di contingenza e selezionare Diploma di laurea dall'elenco a discesa Livello di istruzione.

Figura 5-2
Tavola di contingenza con variabili di strato negli strati della tabella

Tavola di contingenza Income category in thousands * Owns PDA * Level of education

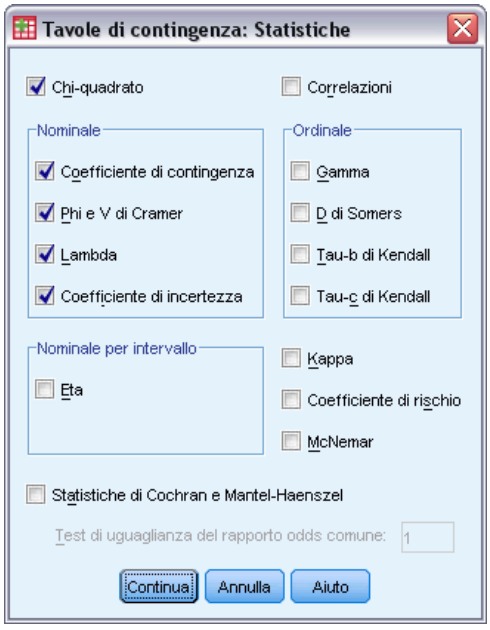
Level of education: College degree

			Owns PDA		
		Statistiche	No	Yes	Totale
Income category in thousands	Under $25	Conteggio	146	50	196
		% entro Owns PDA	15.8%	11.6%	14.5%
	$25 - $49	Conteggio	335	155	490
		% entro Owns PDA	36.3%	35.9%	36.2%
	$50 - $74	Conteggio	187	72	259
		% entro Owns PDA	20.3%	16.7%	19.1%
	$75+	Conteggio	255	155	410
		% entro Owns PDA	27.6%	35.9%	30.3%
Totale		Conteggio	923	432	1355
		% entro Owns PDA	100.0%	100.0%	100.0%

La visualizzazione selezionata della tavola di contingenza mostra le statistiche degli intervistati titolari di una laurea.

Statistiche delle tavole di contingenza

Figura 5-3
Finestra di dialogo Tavole di contingenza: Statistiche



Chi-quadrato. Per tabelle con due righe e due colonne, scegliere Chi-quadrato per calcolare 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 2 × 2, 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 2 × 2 viene calcolato il chi-quadrato corretto di Yates. Per tabelle con un numero qualsiasi di righe e colonne, 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, rho (solo per dati numerici). Il coefficiente rho 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 (nessun ordine intrinseco, ad esempio cattolico, protestante, ebreo), è possibile selezionare il coefficiente Phi e V di Cramér, il Coefficiente di contingenza, Lambda (lambda simmetrico e asimmetrico e tau di Goodman e Kruskal), nonché il Coefficiente di incertezza.

- **Coefficiente di contingenza.** Una misura di associazione basata sul chi-quadrato. Questo coefficiente è sempre compreso tra 0 e 1, dove 0 indica nessuna associazione tra le variabili di riga e colonna e i valori vicini a 1 indicano un alto grado di associazione tra le variabili. Il valore massimo possibile dipende dal numero di righe e di colonne in una tabella.

- **Phi e V di Cramer.** Phi è una misura di associazione calcolata dividendo il chi-quadrato per la dimensione campionaria ed estraendo la radice quadrata del risultato. V di Cramér è una misura di associazione basata sul chi-quadrato.
- **Lambda.** Misura di associazione che riflette la riduzione proporzionale nell'errore quando i valori della variabile indipendente vengono usati per stimare quelli della variabile dipendente. Un valore pari a 1 significa che la variabile indipendente stima perfettamente la variabile dipendente. Un valore pari a 0 significa che la variabile indipendente non è di alcun aiuto nella stima della variabile dipendente.
- **Coefficiente di incertezza.** Misura di associazione che riflette la riduzione proporzionale nell'errore quando i valori di una variabile vengono usati per stimare i valori dell'altra. Un valore di 0,83, ad esempio, indica che la conoscenza di una variabile riduce dell'83% l'errore nella stima dei valori dell'altra variabile. La procedura calcola sia la versione simmetrica, sia quella asimmetrica.

Ordinale. Per tabelle in cui sia le righe che le colonne contengono valori ordinati, 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, selezionare D di Somers.

- **Gamma.** Una misura di associazione simmetrica tra due variabili ordinali che varia tra -1 e 1. I valori prossimi al valore assoluto 1 indicano una forte relazione tra le due variabili. Valori prossimi allo zero indicano scarsità o assenza di relazione. In caso di tabelle a 2 vie verranno visualizzati gamma di ordine zero. Se una tavola di contingenza comprende più di due variabili, verrà calcolato un gamma condizionale per ciascuna sottotabella.
- **D di Somers.** Una misura di associazione tra due variabili ordinali. Varia fra -1 e 1, dove zero indica assenza di associazione e valori prossimi a 1 in valore assoluto indicano forte relazione. È una estensione asimmetrica di gamma dalla quale differisce solo per l'inclusione del numero di coppie non a pari merito nella variabile indipendente. La procedura calcola anche una versione simmetrica di questa statistica.
- **Tau-b di Kendall.** Una misura non parametrica di correlazione per variabili ordinali che tiene conto dei valori pari merito. Il segno del coefficiente indica la direzione della correlazione e il valore assoluto la sua intensità. Valori assoluti maggiori indicano correlazioni maggiori. I valori possibili variano da -1 a +1, ma il valore -1 o +1 può solo essere ottenuto da tabelle quadrate.
- **Tau-c di Kendall.** Misura parametrica di correlazione per variabili ordinali che ignora i valori pari merito. Il segno del coefficiente indica la direzione della correlazione e il valore assoluto la sua intensità. Valori assoluti maggiori indicano correlazioni maggiori. I valori possibili variano da -1 a +1, ma il valore -1 o +1 può solo essere ottenuto da tabelle quadrate.

Nominale per intervallo. Se una variabile è categoriale e l'altra quantitativa, scegliere Eta. La variabile categoriale deve essere codificata numericamente.

- **Eta.** Una misura di associazione che varia fra 0 e 1, dove 0 indica assenza di associazione tra le variabili di riga e colonna e valori prossimi a 1 indicano un grado elevato di associazione. Eta è appropriata per una variabile dipendente misurata su una scala per intervallo e una variabile indipendente con numero limitato di categorie. Vengono calcolati due valori di Eta: il primo assume la variabile di riga, il secondo quella di colonna, come variabile misurata su una scala per intervallo.

Kappa. Il kappa di Cohen misura l'accordo tra due classificatori quando entrambi stanno classificando lo stesso oggetto. Un valore pari a 1 indica accordo perfetto. Un valore pari a 0 indica che l'accordo può essere considerato casuale. Il kappa è basato su una tabella quadrata in cui i valori di riga e di colonna rappresentano la stessa scala. A ciascuna cella contenente valori osservati per una variabile ma non per l'altra viene assegnato un conteggio di 0. Il kappa non viene calcolato se il tipo di archiviazione dati (stringa o numerico) non è uguale per le due variabili. Nel caso della variabile stringa, entrambe le variabili devono avere la stessa lunghezza definita.

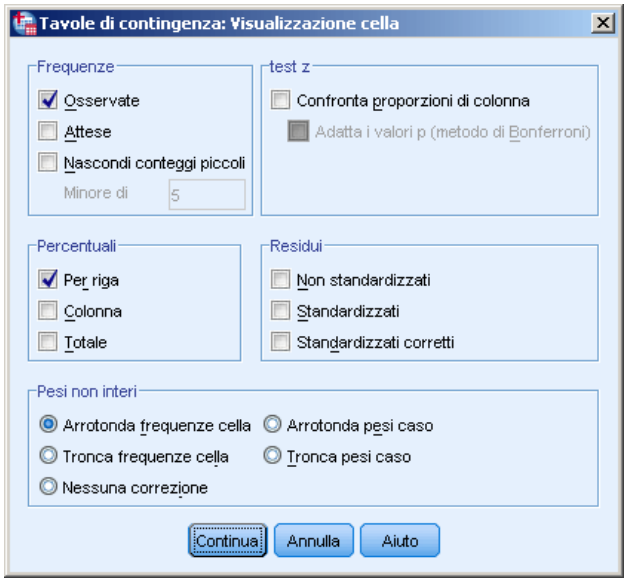
Rischio. Nelle tabelle 2x2 il rischio relativo misura l'intensità dell'associazione tra la presenza di un fattore e la manifestazione di un evento. Un valore pari a 1 indica che il fattore non è associato all'evento. Il rapporto "odds ratio" può essere usato come stima del fattore di rischio quando la presenza del fattore è rara.

McNemar. Test non parametrico per due variabili dicotomiche correlate. Verifica la presenza di cambiamenti nelle risposte mediante la distribuzione chi-quadrato. Il test è molto utile in disegni sperimentali del tipo 'prima e dopo', per rilevare cambiamenti di risposta. Per tabelle quadrate di dimensioni maggiori, viene calcolato il test della simmetria di McNemar-Bowker.

Statistiche di Cochran e Mantel-Haenszel. Le statistiche di Cochran e Mantel-Haenszel possono essere usate come test di indipendenza fra un fattore binario e una variabile di risposta binaria. Le statistiche sono corrette per modelli covariati definiti da una o più variabili di controllo. Si noti che mentre altre statistiche vengono calcolate strato per strato, le statistiche di Cochran e Mantel-Haenszel vengono calcolate per tutti i livelli.

Visualizzazione delle celle delle tavole di contingenza

Figura 5-4
Finestra di dialogo Tavole di contingenza: Visualizzazione cella



Per facilitare l'individuazione di modelli di dati che danno origine a un test chi-quadrato significativo, la procedura per le tavole di contingenza visualizza le frequenze attese e tre tipi di residui (devianze) che misurano la differenza tra le frequenze osservate e quelle attese. Ogni cella

della tabella può contenere qualsiasi combinazione dei conteggi, delle percentuali e dei residui selezionati.

Conteggi. Il numero di casi effettivamente osservati e il numero di casi attesi se le variabili di riga e di colonna sono reciprocamente indipendenti. È possibile scegliere di nascondere i conteggi inferiori a un numero intero specificato. I valori nascosti verranno visualizzati come <N, dove N è il numero intero specificato. Il numero intero specificato deve essere maggiore o uguale a 2; è consentito anche il valore 0 per indicare che nessun conteggio deve essere nascosto.

Confronta proporzioni di colonna. Questa opzione calcola confronti a coppie delle proporzioni di colonna e indica le coppie di colonne (di una determinata riga) significativamente diverse. Le differenze significative vengono indicate nella tabella tavola di contingenza con una formattazione tipo APA mediante lettere in formato pedice e vengono calcolate al livello di significatività di 0,05. *Nota*: se questa opzione viene specificata senza selezionare i conteggi osservati o le percentuali di colonna, i conteggi osservati sono inclusi nella tabella delle tavole di contingenza, con lettere di stile APA in formato pedice a indicare i risultati dei test delle proporzioni di colonna.

- **Adatta i valori p per confronti multipli (metodo Bonferroni).** I confronti a coppie delle proporzioni di colonna utilizzano la correzione Bonferroni, che adatta il livello di significatività osservato in base al fatto che vengono eseguiti confronti multipli.

Percentuali. Le percentuali possono essere aggiunte nelle righe o nelle colonne. Sono disponibili anche le percentuali del numero totale di casi rappresentati nella tabella (a strato unico). *Nota*: se nel gruppo Conteggi viene selezionata l'opzione Nascondi conteggi piccoli, vengono nascoste anche le percentuali associate ai conteggi nascosti.

Residui. I residui semplici non standardizzati forniscono la differenza tra valori osservati e attesi. Sono inoltre disponibili residui standardizzati e standardizzati corretti.

- **Non standardizzati.** La differenza fra un valore osservato e un valore previsto. Per valore atteso si intendo il numero di casi atteso nella cella in assenza di relazione tra le due variabili. Un residuo positivo indica che ci sono più casi nella cella di quanti ce ne sarebbero se le variabili di riga e di colonna fossero indipendenti.
- **Standardizzati.** Il residuo diviso per una stima della deviazione standard. Il residuo standardizzato, conosciuto anche come residuo di Pearson, ha media 0 e deviazione standard 1.
- **Standardizzati corretti.** Il residuo di una cella (valore osservato meno valore atteso) diviso per una stima del suo errore standard. Viene espresso in unità di deviazione standard sopra o sotto la media.

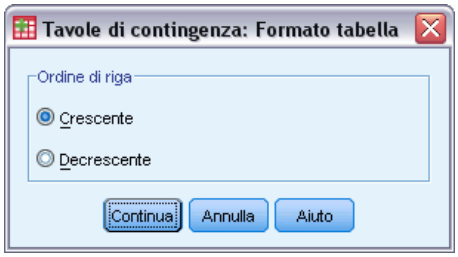
Pesi non interi. I conteggi di cella in genere sono valori interi, in quanto rappresentano il numero di casi in ogni cella. Se, tuttavia, il file di dati è attualmente ponderato in base a una variabile di ponderazione con valori frazionari (ad esempio, 1,25), i conteggi di cella possono essere espressi anche in valori frazionari. È possibile troncare o arrotondare i valori prima o dopo aver calcolato i conteggi di cella oppure utilizzare conteggi di cella frazionari per la visualizzazione delle tabelle e dei calcoli statistici.

- **Arrotonda frequenze cella.** I pesi dei casi vengono usati come tali, mentre i pesi accumulati nelle celle vengono arrotondati prima di calcolare qualsiasi statistica.
- **Tronca conteggi cella.** I pesi dei casi sono usati come tali, ma i pesi accumulati nelle celle sono troncati prima di calcolare qualunque statistica.
- **Arrotonda pesi caso.** I pesi dei casi vengono arrotondati prima dell'uso.

- **Tronca pesi caso.** I pesi dei casi sono troncati prima dell'uso.
- **Nessuna correzione.** I pesi di caso vengono utilizzati come sono e vengono utilizzati anche i conteggi di cella frazionari. Quando tuttavia è richiesto l'utilizzo dell'opzione Statistiche esatte (disponibile solo con l'opzione Test esatti), i pesi di caso delle celle verranno troncati o arrotondati prima del calcolo delle statistiche esatte.

Formato tabella delle tavole di contingenza

Figura 5-5
Finestra di dialogo Tavole di contingenza: Formato tabella



È possibile disporre le righe nell'ordine crescente o decrescente dei valori della variabile di riga.

Capitolo
6

Riassumi

La procedura Riassumi consente di calcolare le statistiche di sottogruppo per le variabili all'interno delle categorie di una o più variabili di raggruppamento. Tutti i livelli della variabile di raggruppamento vengono incrociati. È possibile scegliere l'ordine in cui vengono visualizzate le statistiche. Per ciascuna variabile di tutte le categorie verranno inoltre visualizzate le statistiche riassuntive. I valori di ciascuna categoria possono essere inseriti nell'elenco o eliminati, ma negli insiemi di dati di grandi dimensioni, è possibile scegliere di elencare solo i primi *n* casi.

Esempio. Qual è l'importo medio delle vendite per area e industria del cliente? Si potrebbe scoprire che l'importo medio delle vendite è leggermente superiore nell'area occidentale rispetto alle altre aree e che ai clienti di quest'area è associato l'importo medio più alto.

Statistiche. Somma, numero di casi, media, mediana, mediana dei gruppi, errore standard della media, minimo, massimo, intervallo, valore della prima categoria della variabile di raggruppamento, valore dell'ultima categoria della variabile di raggruppamento, deviazione standard, varianza, curtosi, errore standard della curtosi, asimmetria, errore standard dell'asimmetria, percentuale della somma totale, percentuale del conteggio totale, percentuale della somma di gruppo, percentuale dei conteggi di gruppo, media geometrica, media armonica.

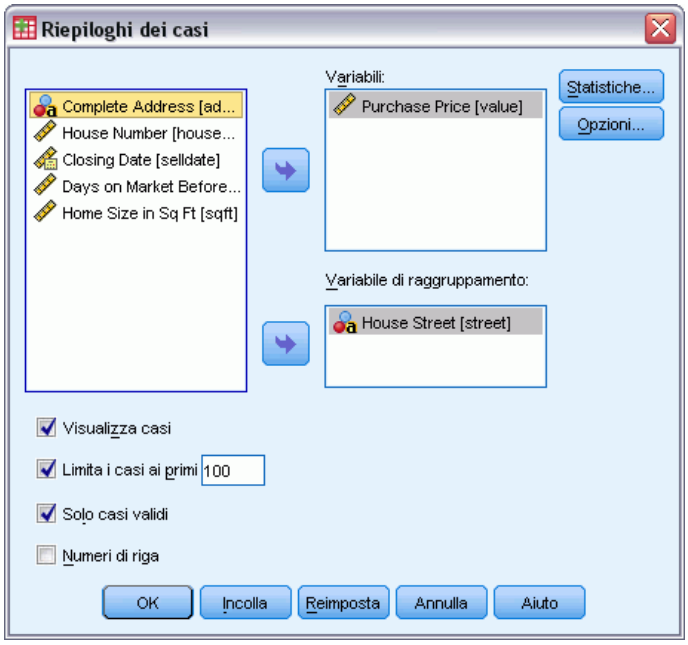
Dati. Le variabili di raggruppamento sono variabili categoriali che possono contenere valori stringa o numerici. Il numero di categorie dovrebbe essere limitato. Le altre variabili dovrebbero essere classificabili.

Assunzioni. Alcune delle statistiche di sottogruppo facoltative, quali la media e la deviazione standard, sono basate sulla teoria della normalità e sono idonee per le variabili quantitative con distribuzione simmetrica dei dati. La mediana e l'intervallo sono statistiche robuste, idonee per le variabili quantitative che possono o meno soddisfare l'assunzione di normalità.

Per ottenere riepiloghi dei casi

- Dai menu, scegliere:
  Analizza > Report > Riepiloghi dei casi...

Figura 6-1
Finestra di dialogo Riepiloghi dei casi



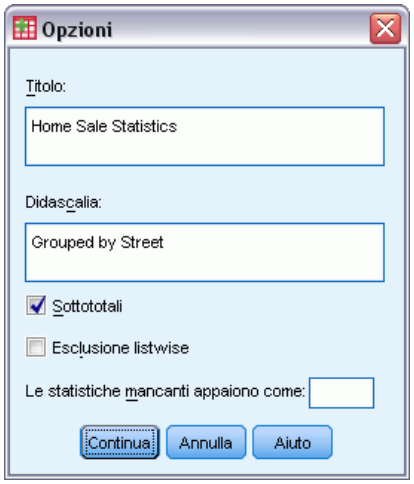
▶ Selezionare una o più variabili.

Se lo si desidera, è possibile:

- Selezionare una o più variabili di raggruppamento per suddividere i dati in sottogruppi.
- Fare clic su Opzioni per modificare il titolo dell'output, aggiungere una didascalia al di sotto dell'output o escludere casi con valori mancanti.
- Fare clic su Statistiche per visualizzare statistiche facoltative.
- Selezionare Visualizza casi per visualizzare un elenco dei casi inclusi in ciascun sottogruppo. Per impostazione predefinita, vengono elencati solo i primi 100 casi nel file. È possibile aumentare o ridurre il valore di Limita i casi ai primi*n* oppure deselezionare l'opzione per visualizzare l'elenco di tutti i casi.

Riassumi: Opzioni

Figura 6-2
Finestra di dialogo Opzioni

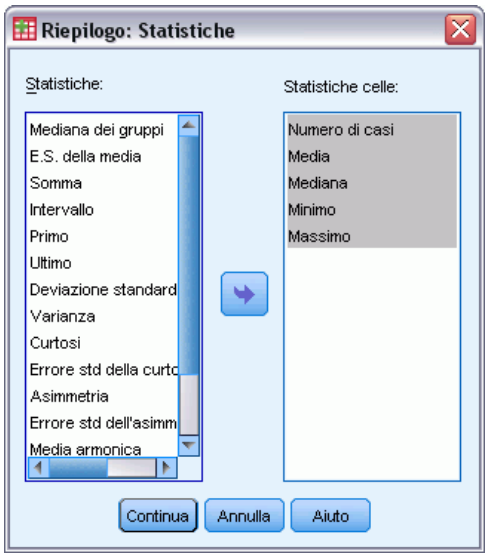


SPSS consente di modificare il titolo dell'output o di aggiungere una didascalia che verrà visualizzata sotto alla tabella di output. È possibile controllare gli a capo automatici dei titoli e delle didascalie digitando \n dove si desidera inserire un'interruzione di riga nel testo.

È inoltre possibile scegliere di visualizzare o eliminare i sottototali e di includere o escludere casi con valori mancanti per qualsiasi variabile utilizzata nelle analisi. È spesso consigliabile contrassegnare nell'output i casi mancanti utilizzando un punto o un asterisco. Immettere un carattere, una frase o codice che si desidera venga visualizzato per indicare che un valore è mancante. In caso contrario, ai casi mancanti non verrà applicato alcun identificatore nell'output.

Riassumi: Statistiche

Figura 6-3
Finestra di dialogo delle statistiche dei report riassuntive



È possibile scegliere una o più delle seguenti statistiche di sottogruppo per le variabili all'interno di ogni categoria di ciascuna variabile di raggruppamento: somma, numero di casi, media, mediana, mediana dei gruppi, errore standard della media, minimo, massimo, intervallo, valore della prima categoria della variabile di raggruppamento, valore dell'ultima categoria della variabile di raggruppamento, deviazione standard, varianza, curtosi, errore standard della curtosi, asimmetria, errore standard dell'asimmetria, percentuale della somma totale, percentuale del conteggio totale, percentuale della somma di gruppo, percentuale dei conteggi di gruppo, media geometrica, media armonica. L'ordine in cui compaiono le statistiche nell'elenco Statistiche di cella corrisponde all'ordine in cui verranno visualizzate nell'output. Per ciascuna variabile vengono visualizzate anche le statistiche riassuntive in tutte le categorie.

Primo. Visualizza il primo valore incontrato nel file di dati.

Media geometrica. La radice ennesima del prodotto dei valori, dove n è il numero di casi.

Mediana dei gruppi. La mediana calcolata su valori che rappresentano gruppi. Ad esempio la mediana della fascia di età.

Media armonica. Usata per stimare una dimensione media dei gruppi quando le dimensioni campionarie dei gruppi non sono uguali. La media armonica è il numero di campioni diviso per la somma dei reciproci delle dimensioni campionarie.

Curtosi. Una misura di quanto le osservazioni si trovino raggruppate nelle code. Per la distribuzione normale, il valore della statistica di curtosi è zero. Una curtosi positiva indica che, rispetto a una distribuzione normale, le osservazioni sono più raggruppate intorno al centro della distribuzione e hanno code più sottili fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione leptocurtica sono più spesse rispetto a una distribuzione normale. Una curtosi negativa indica che, rispetto a una distribuzione normale, le osservazioni sono meno

raggruppate e hanno code più spesse fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione platicurtica sono più sottili rispetto a una distribuzione normale.

Ultimo. Visualizza l'ultimo valore di dati incontrato nel file di dati.

Massimo. Il valore più alto di una variabile numerica.

Media. Una misura di tendenza centrale. La somma dei valori di tutte le osservazioni divisa per il numero di osservazioni. Viene anche detta media aritmetica.

Mediana. È il valore sopra il quale e sotto il quale ricade la metà dei casi, il 50-esimo percentile. Se il numero di casi è pari, la mediana è pari alla media dei due casi centrali quando questi sono ordinati secondo l'ordine ascendente o discendente. La mediana è una misura di tendenza centrale non sensibile ai valori anomali, a differenza della media che può essere influenzata da valori eccezionalmente bassi o alti.

Minimo. Il valore più basso assunto da una variabile numerica.

N. Il numero di casi (osservazioni o record).

% del numero di casi totale. Percentuale del numero di casi totale in ogni categoria.

% della somma totale. Percentuale della somma totale in ogni categoria.

Intervallo. La differenza tra il valore massimo ed il valore minimo di una variabile numerica.

Asimmetria. Una misura dell'asimmetria di una distribuzione. La distribuzione normale è simmetrica e ha un valore di asimmetria pari a 0. Una distribuzione con una notevole asimmetria positiva ha una lunga coda a destra. Una distribuzione con asimmetria negativa ha una coda a sinistra. In generale un'asimmetria con valore più che doppio dell'errore standard indica lo scostamento dalla normale simmetria.

Errore standard della curtosi. L'ipotesi di normalità può essere rifiutata se questo rapporto è maggiore di 2 in valore assoluto. Un valore positivo elevato per la curtosi indica che le code della distribuzione sono più lunghe di quelle di una distribuzione normale; un valore negativo per la curtosi indica code più corte, simili a quelle di una distribuzione uniforme a forma di scatola.

Errore standard dell'asimmetria. Il rapporto fra l'asimmetria di una distribuzione e il suo errore standard viene usato come test di normalità. L'ipotesi di normalità può essere rifiutata se questo rapporto è maggiore di 2 in valore assoluto. Un valore positivo elevato per l'asimmetria indicata una coda a destra lunga; un valore negativo estremo indica una coda a sinistra lunga.

Somma. La somma o il totale di tutti i valori non mancanti di tutti i casi.

Varianza. Una misura della dispersione dei valori intorno alla media. È calcolata come somma dei quadrati degli scostamenti dalla media, divisa per il numero totale delle osservazioni valide meno 1. La varianza è espressa in quadrati dell'unità di misura della variabile.

Capitolo
7

Medie

La procedura Medie consente di calcolare le medie dei sottogruppi e le statistiche univariate correlate per le variabili dipendenti all'interno delle categorie di una o più variabili indipendenti. È inoltre possibile ottenere analisi univariate della varianza, eta e test di linearità.

Esempio. Misurare la quantità media di grasso assorbita da tre diversi tipi di olii alimentari ed eseguire un'analisi univariata della varianza per verificare se le medie differiscono.

Statistiche. Somma, numero di casi, media, mediana, mediana dei gruppi, errore standard della media, minimo, massimo, intervallo, valore della prima categoria della variabile di raggruppamento, valore dell'ultima categoria della variabile di raggruppamento, deviazione standard, varianza, curtosi, errore standard della curtosi, asimmetria, errore standard dell'asimmetria, percentuale della somma totale, percentuale del conteggio totale, percentuale della somma di gruppo, percentuale dei conteggi di gruppo, media geometrica, media armonica. Le opzioni includono analisi della varianza, eta, eta quadrato, e test di linearità R e R^2.

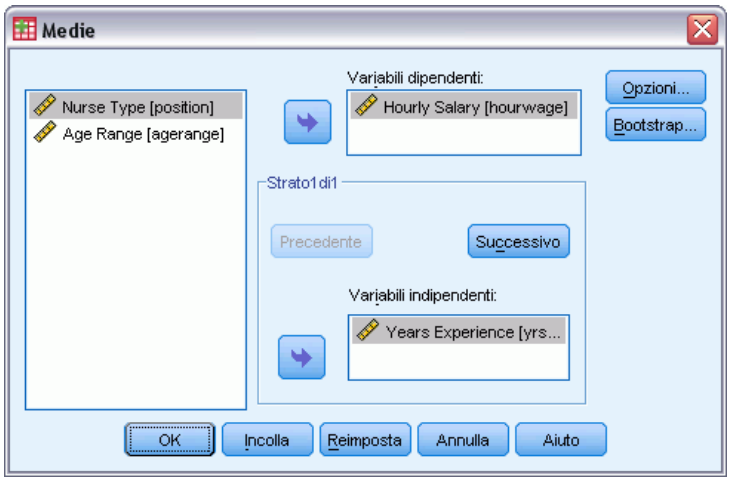
Dati. Le variabili dipendenti sono quantitative e le variabili indipendenti sono categoriali. I valori delle variabili categoriali possono essere di tipo numerico o stringa.

Assunzioni. Alcune delle statistiche di sottogruppo facoltative, quali la media e la deviazione standard, sono basate sulla teoria della normalità e sono idonee per le variabili quantitative con distribuzione simmetrica dei dati. La mediana è una statistica robusta, idonea per le variabili quantitative che possono o meno soddisfare l'ipotesi di normalità. L'analisi della varianza è robusta per quanto riguarda le alterazioni della normalità, ma i dati in ciascuna cella devono essere simmetrici. L'analisi della varianza assume inoltre che i gruppi provengano da popolazioni con valori di varianza uguali. Per verificare questa ipotesi, utilizzare il test di omogeneità della varianza di Levene, disponibile nella procedura ANOVA univariata.

Per ottenere le medie dei sottogruppi

- Dai menu, scegliere:
  Analizza > Confronta medie > Medie...

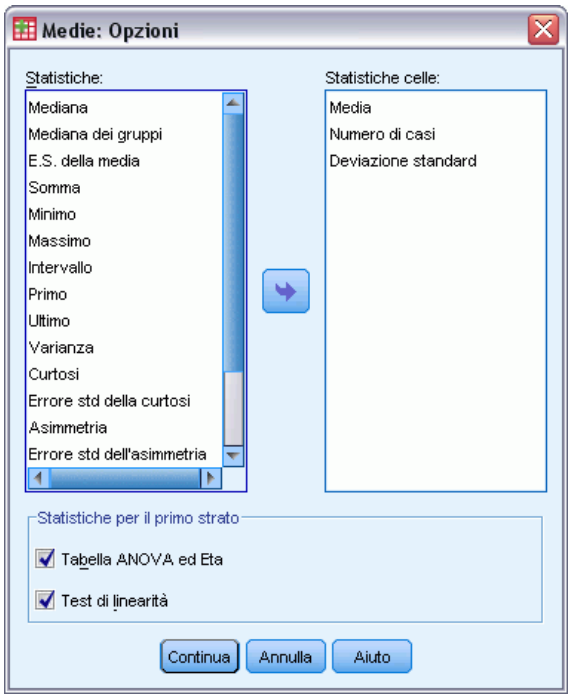
Figura 7-1
Finestra di dialogo Medie



- Selezionare una o più variabili dipendenti.
- Usare uno dei seguenti metodi per selezionare le variabili categoriali indipendenti:
  - Selezionare una o più variabili indipendenti. Per ciascuna variabile indipendente vengono visualizzati risultati distinti.
  - Selezionare uno o più strati di variabili indipendenti. Ogni strato suddivide ulteriormente il campione. Se è presente una sola variabile indipendente nello Strato 1 e una sola nello Strato 2, i risultati verranno visualizzati in una tabella incrociata e non in tabelle distinte per ciascuna variabile indipendente.
- Oppure fare clic su Opzioni per ottenere statistiche facoltative, analisi della tabella di varianza, età, età quadrato, R e R^2.

Medie: Opzioni

Figura 7-2
Finestra di dialogo Medie: Opzioni



È possibile scegliere una o più delle seguenti statistiche di sottogruppo per le variabili all'interno di ogni categoria di ciascuna variabile di raggruppamento: somma, numero di casi, media, mediana, mediana dei gruppi, errore standard della media, minimo, massimo, intervallo, valore della prima categoria della variabile di raggruppamento, valore dell'ultima categoria della variabile di raggruppamento, deviazione standard, varianza, curtosi, errore standard della curtosi, asimmetria, errore standard dell'asimmetria, percentuale della somma totale, percentuale del numero*N* totale, percentuale della somma, percentuale del *numero di casi* in, media geometrica, media armonica. È possibile modificare l'ordine in cui compaiono le statistiche per i sottogruppi. L'ordine in cui compaiono le statistiche nella lista Statistiche di cella corrisponde all'ordine in cui verranno visualizzate nell'output. Per ciascuna variabile vengono visualizzate anche le statistiche riassuntive in tutte le categorie.

Primo. Visualizza il primo valore incontrato nel file di dati.

Media geometrica. La radice ennesima del prodotto dei valori, dove n è il numero di casi.

Mediana dei gruppi. La mediana calcolata su valori che rappresentano gruppi. Ad esempio la mediana della fascia di età.

Media armonica. Usata per stimare una dimensione media dei gruppi quando le dimensioni campionarie dei gruppi non sono uguali. La media armonica è il numero di campioni diviso per la somma dei reciproci delle dimensioni campionarie.

Curtosi. Una misura di quanto le osservazioni si trovino raggruppate nelle code. Per la distribuzione normale, il valore della statistica di curtosi è zero. Una curtosi positiva indica che, rispetto a una distribuzione normale, le osservazioni sono più raggruppate intorno al centro della distribuzione e hanno code più sottili fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione leptocurtica sono più spesse rispetto a una distribuzione normale. Una curtosi negativa indica che, rispetto a una distribuzione normale, le osservazioni sono meno raggruppate e hanno code più spesse fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione platicurtica sono più sottili rispetto a una distribuzione normale.

Ultimo. Visualizza l'ultimo valore di dati incontrato nel file di dati.

Massimo. Il valore più alto di una variabile numerica.

Media. Una misura di tendenza centrale. La somma dei valori di tutte le osservazioni divisa per il numero di osservazioni. Viene anche detta media aritmetica.

Mediana. È il valore sopra il quale e sotto il quale ricade la metà dei casi, il 50-esimo percentile. Se il numero di casi è pari, la mediana è pari alla media dei due casi centrali quando questi sono ordinati secondo l'ordine ascendente o discendente. La mediana è una misura di tendenza centrale non sensibile ai valori anomali, a differenza della media che può essere influenzata da valori eccezionalmente bassi o alti.

Minimo. Il valore più basso assunto da una variabile numerica.

N. Il numero di casi (osservazioni o record).

Percentuale del numero di casi totale. Percentuale del numero di casi totale in ogni categoria.

Percentuale della somma totale. Percentuale della somma totale in ogni categoria.

Intervallo. La differenza tra il valore massimo ed il valore minimo di una variabile numerica.

Asimmetria. Una misura dell'asimmetria di una distribuzione. La distribuzione normale è simmetrica e ha un valore di asimmetria pari a 0. Una distribuzione con una notevole asimmetria positiva ha una lunga coda a destra. Una distribuzione con asimmetria negativa ha una coda a sinistra. In generale un'asimmetria con valore più che doppio dell'errore standard indica lo scostamento dalla normale simmetria.

Errore standard della curtosi. L'ipotesi di normalità può essere rifiutata se questo rapporto è maggiore di 2 in valore assoluto. Un valore positivo elevato per la curtosi indica che le code della distribuzione sono più lunghe di quelle di una distribuzione normale; un valore negativo per la curtosi indica code più corte, simili a quelle di una distribuzione uniforme a forma di scatola.

Errore standard dell'asimmetria. Il rapporto fra l'asimmetria di una distribuzione e il suo errore standard viene usato come test di normalità. L'ipotesi di normalità può essere rifiutata se questo rapporto è maggiore di 2 in valore assoluto. Un valore positivo elevato per l'asimmetria indicata una coda a destra lunga; un valore negativo estremo indica una coda a sinistra lunga.

Somma. La somma o il totale di tutti i valori non mancanti di tutti i casi.

Varianza. Una misura della dispersione dei valori intorno alla media. È calcolata come somma dei quadrati degli scostamenti dalla media, divisa per il numero totale delle osservazioni valide meno 1. La varianza è espressa in quadrati dell'unità di misura della variabile.

Statistiche per il primo strato

Tabella ANOVA ed Eta. Visualizza una tabella di analisi della varianza univariata e calcola le misure di associazione eta ed eta quadrato per ogni variabile indipendente del primo strato.

Test di linearità. Calcola la somma dei quadrati, i gradi di libertà e la media dei quadrati associata alle componenti lineari e non lineari, nonché il rapporto F, R e R-quadrato. La linearità non viene calcolata se la variabile indipendente è una stringa corta.

Capitolo
8

Cubi OLAP

La procedura Cubi OLAP (Online Analytical Processing) consente di calcolare i totali, le medie e le altre statistiche univariate per le variabili riassunte continue all'interno delle categorie di una o più variabili di raggruppamento categoriali. Nella tabella viene creato uno strato distinto per ciascuna categoria di ogni variabile di raggruppamento.

Esempio. Le vendite totali e medie di diverse aree e le linee di prodotti all'interno delle aree.

Statistiche. Somma, numero di casi, media, mediana, mediana dei gruppi, errore standard della media, minimo, massimo, intervallo, valore della prima categoria della variabile di raggruppamento, valore dell'ultima categoria della variabile di raggruppamento, deviazione standard, varianza, curtosi, errore standard della curtosi, asimmetria, errore standard dell'asimmetria, percentuale della somma totale, percentuale del numero di casi totale, percentuale della somma totale entro variabili di raggruppamento, percentuale del numero di casi totale entro variabili di raggruppamento, media geometrica, media armonica.

Dati. Le variabili riassunte sono quantitative (variabili continue misurate su una scala di intervallo o di rapporto) e le variabili di raggruppamento sono categoriali. I valori delle variabili categoriali possono essere di tipo numerico o stringa.

Assunzioni. Alcune delle statistiche di sottogruppo facoltative, quali la media e la deviazione standard, sono basate sulla teoria della normalità e sono idonee per le variabili quantitative con distribuzione simmetrica dei dati. La mediana e l'intervallo sono statistiche robuste, idonee per le variabili quantitative che possono o meno soddisfare l'ipotesi di normalità.

Per ottenere cubi OLAP

▶ Dai menu, scegliere:
Analizza > Report > Cubi OLAP...

Figura 8-1
Finestra di dialogo Cubi OLAP

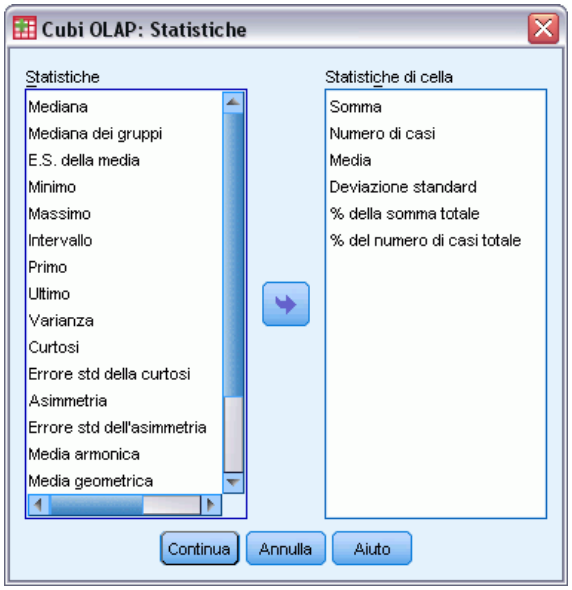
- Selezionare una o più variabili riassunte continue.
- Selezionare una o più variabili categoriali di raggruppamento

Oppure:

- Selezionare statistiche riassuntive diverse (fare clic su Statistiche). Prima di selezionare le statistiche riassuntive è necessario selezionare una o più variabili di raggruppamento.
- Calcolare differenze tra coppie di variabili e coppie di gruppi definiti da una variabile di raggruppamento (fare clic su Differenze).
- Creare titoli di tabella personalizzati (fare clic su Titolo).
- Nascondere i conteggi inferiori a un numero intero specificato. I valori nascosti verranno visualizzati come <N, dove N è il numero intero specificato. Il numero intero specificato deve essere maggiore o uguale a 2.

Cubi OLAP: Statistiche

Figura 8-2
Finestra di dialogo Cubi OLAP: Statistiche



È possibile scegliere una o più delle seguenti statistiche di sottogruppo per le variabili riassunte all'interno di ogni categoria di ciascuna variabile di raggruppamento: Somma, numero di casi, media, mediana, mediana dei gruppi, errore standard della media, minimo, massimo, intervallo, valore della prima categoria della variabile di raggruppamento, valore dell'ultima categoria della variabile di raggruppamento, deviazione standard, varianza, curtosi, errore standard della curtosi, asimmetria, errore standard dell'asimmetria, percentuale della somma totale, percentuale del numero di casi totale, percentuale della somma totale entro variabili di raggruppamento, percentuale del numero di casi totale entro variabili di raggruppamento, media geometrica, media armonica.

È possibile modificare l'ordine in cui compaiono le statistiche per i sottogruppi. L'ordine in cui compaiono le statistiche nella lista Statistiche di cella corrisponde all'ordine in cui verranno visualizzate nell'output. Per ciascuna variabile vengono visualizzate anche le statistiche riassuntive in tutte le categorie.

Primo. Visualizza il primo valore incontrato nel file di dati.

Media geometrica. La radice ennesima del prodotto dei valori, dove n è il numero di casi.

Mediana dei gruppi. La mediana calcolata su valori che rappresentano gruppi. Ad esempio la mediana della fascia di età.

Media armonica. Usata per stimare una dimensione media dei gruppi quando le dimensioni campionarie dei gruppi non sono uguali. La media armonica è il numero di campioni diviso per la somma dei reciproci delle dimensioni campionarie.

Curtosi. Una misura di quanto le osservazioni si trovino raggruppate nelle code. Per la distribuzione normale, il valore della statistica di curtosi è zero. Una curtosi positiva indica che, rispetto a una distribuzione normale, le osservazioni sono più raggruppate intorno al centro della distribuzione e hanno code più sottili fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione leptocurtica sono più spesse rispetto a una distribuzione normale. Una curtosi negativa indica che, rispetto a una distribuzione normale, le osservazioni sono meno raggruppate e hanno code più spesse fino ai valori estremi della distribuzione; a quel punto, le code della distribuzione platicurtica sono più sottili rispetto a una distribuzione normale.

Ultimo. Visualizza l'ultimo valore di dati incontrato nel file di dati.

Massimo. Il valore più alto di una variabile numerica.

Media. Una misura di tendenza centrale. La somma dei valori di tutte le osservazioni divisa per il numero di osservazioni. Viene anche detta media aritmetica.

Mediana. È il valore sopra il quale e sotto il quale ricade la metà dei casi, il 50-esimo percentile. Se il numero di casi è pari, la mediana è pari alla media dei due casi centrali quando questi sono ordinati secondo l'ordine ascendente o discendente. La mediana è una misura di tendenza centrale non sensibile ai valori anomali, a differenza della media che può essere influenzata da valori eccezionalmente bassi o alti.

Minimo. Il valore più basso assunto da una variabile numerica.

N. Il numero di casi (osservazioni o record).

Percentuale del numero di casi in. Percentuale del numero di casi di una variabile di raggruppamento entro le categorie di altre variabili di raggruppamento. Se esiste una sola variabile di raggruppamento, questo valore è uguale alla percentuale del conteggio totale

% della somma in. Percentuale della somma di una variabile di raggruppamento entro le categorie di altre variabili di raggruppamento. Se esiste una sola variabile di raggruppamento, questo valore è uguale alla percentuale della somma totale.

% del numero di casi totale. Percentuale del numero di casi totale in ogni categoria.

% della somma totale. Percentuale della somma totale in ogni categoria.

Intervallo. La differenza tra il valore massimo ed il valore minimo di una variabile numerica.

Asimmetria. Una misura dell'asimmetria di una distribuzione. La distribuzione normale è simmetrica e ha un valore di asimmetria pari a 0. Una distribuzione con una notevole asimmetria positiva ha una lunga coda a destra. Una distribuzione con asimmetria negativa ha una coda a sinistra. In generale un'asimmetria con valore più che doppio dell'errore standard indica lo scostamento dalla normale simmetria.

Errore standard della curtosi. L'ipotesi di normalità può essere rifiutata se questo rapporto è maggiore di 2 in valore assoluto. Un valore positivo elevato per la curtosi indica che le code della distribuzione sono più lunghe di quelle di una distribuzione normale; un valore negativo per la curtosi indica code più corte, simili a quelle di una distribuzione uniforme a forma di scatola.

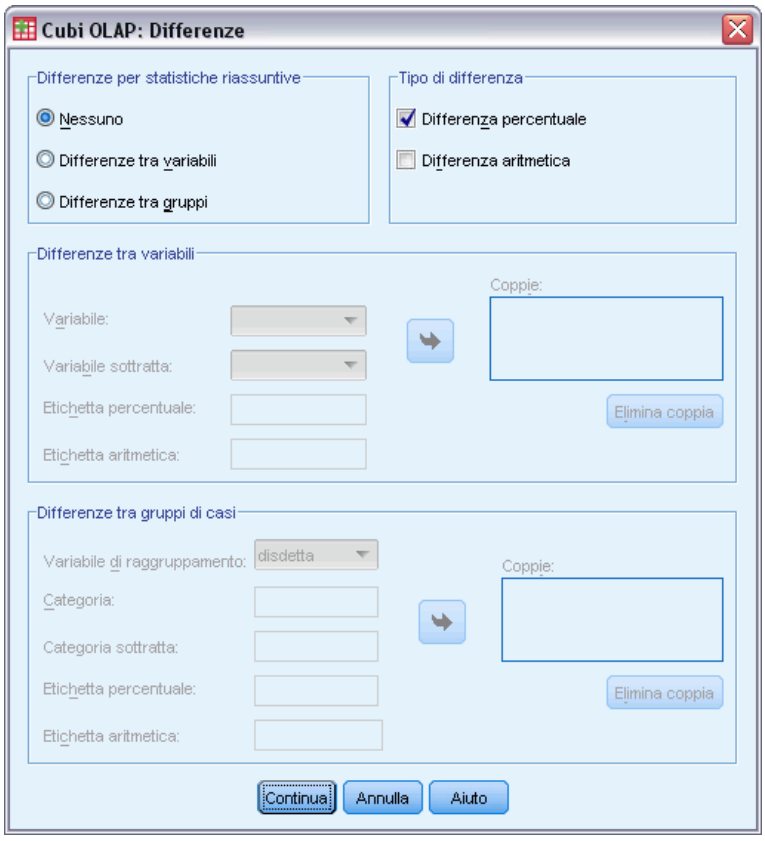
Errore standard dell'asimmetria. Il rapporto fra l'asimmetria di una distribuzione e il suo errore standard viene usato come test di normalità. L'ipotesi di normalità può essere rifiutata se questo rapporto è maggiore di 2 in valore assoluto. Un valore positivo elevato per l'asimmetria indicata una coda a destra lunga; un valore negativo estremo indica una coda a sinistra lunga.

Somma. La somma o il totale di tutti i valori non mancanti di tutti i casi.

Varianza. Una misura della dispersione dei valori intorno alla media. È calcolata come somma dei quadrati degli scostamenti dalla media, divisa per il numero totale delle osservazioni valide meno 1. La varianza è espressa in quadrati dell'unità di misura della variabile.

Cubi OLAP: Differenze

Figura 8-3
Finestra di dialogo Cubi OLAP: Differenze



Questa finestra di dialogo consente di calcolare le differenze aritmetiche e percentuali tra variabili riassunte o tra gruppi definiti da una variabile di raggruppamento. Le differenze vengono calcolate per tutte le misure selezionate nella finestra di dialogo Cubi OLAP: Statistiche.

Differenze tra variabili. Consente di calcolare le differenze tra coppie di variabili. I valori delle statistiche riassuntive della seconda variabile di ogni coppia (la variabile sottratta) vengono sottratti dai valori delle statistiche riassuntive della prima variabile della coppia. Per le differenze percentuali, il valore della variabile riassunta della variabile sottratta viene utilizzato al denominatore. È necessario selezionare almeno due variabili riassunte nella finestra di dialogo principale prima di specificare le differenze tra variabili.

Differenze tra gruppi di casi. Consente di calcolare le differenze tra coppie di gruppi definiti da una variabile di raggruppamento. I valori delle statistiche riassuntive della seconda categoria di ogni coppia (la categoria sottratta) vengono sottratti dai valori delle statistiche riassuntive della prima categoria della coppia. Per le differenze percentuali, il valore delle statistiche riassuntive della categoria sottratta viene utilizzato al denominatore. È necessario selezionare una o più variabili di raggruppamento nella finestra di dialogo principale prima di specificare le differenze tra gruppi.

Cubi OLAP: Titolo

Figura 8-4
Finestra di dialogo Cubi OLAP: Titolo

È possibile modificare il titolo dell'output o aggiungere una didascalia che verrà visualizzata sotto la tabella dell'output. È inoltre possibile impostare gli a capo automatici dei titoli e delle didascalie digitando \n dove si desidera inserire un'interruzione di riga nel testo.

Capitolo
9

Test T

Sono disponibili tre tipi di test *t*:

Test T per campioni indipendenti (test T per due campioni). Consente di confrontare le medie di una variabile per due gruppi di casi. Vengono fornite statistiche descrittive per ciascun gruppo, il test di Levene di uguaglianza delle varianze, i valori *t* di uguaglianza e non uguaglianza della varianza e un intervallo di confidenza al 95% per la differenza tra le medie.

Test T per campioni appaiati (test T dipendente). Consente di confrontare le medie di due variabili per un singolo gruppo. Questo test viene utilizzato anche per disegni relativi a studi di confronti tra coppie o di casi di controllo. Vengono fornite statistiche descrittive per le variabili oggetto del test, la correlazione tra di esse, le statistiche descrittive per le differenze appaiate, il test *t* e un intervallo di confidenza al 95%.

Test T per un campione. Consente di confrontare la media di una variabile con un valore noto o un valore ipotizzato. Con il test *t* vengono visualizzate anche le statistiche descrittive per le variabili oggetto del test. L'output predefinito include un intervallo di confidenza al 95% per la differenza tra la media della variabile oggetto del test e il valore ipotizzato per il test.

T per campioni indipendenti

Il test T per campioni indipendenti consente di confrontare le medie relative a due gruppi di casi. Nel test, i soggetti dovrebbero essere assegnati in modo casuale a due gruppi. In questo modo, le eventuali differenze nella riposta saranno dovute alla modalità di elaborazione (o alla mancata elaborazione) e non ad altri fattori. Ciò non si verifica se si esegue il confronto tra il reddito medio di soggetti maschili e femminili. Non è infatti possibile assegnare in modo casuale una persona al sesso maschile o femminile. In questi casi, è necessario assicurarsi che le differenze relative ad altri fattori non comportino un mascheramento o l'incremento di differenze significative nelle medie. Le differenze nel reddito medio possono essere influenzate da fattori quali il livello di educazione e non solo dal sesso al quale appartengono i soggetti.

Esempio. I pazienti con pressione sanguigna alta vengono assegnati in modo casuale a un gruppo di controllo e a un gruppo di trattamento. Ai soggetti del gruppo di controllo vengono somministrate medicine innocue e ai soggetti del gruppo trattato viene somministrato un nuovo farmaco che si ritiene possa far diminuire la pressione sanguigna. Al termine di un trattamento di due mesi, viene utilizzato il test *t* per due campioni allo scopo di confrontare i valori medi della pressione sanguigna nel gruppo di controllo e nel gruppo trattato. La pressione di ogni paziente viene misurata una volta e ciascun paziente appartiene a un solo gruppo.

Statistiche. Per ogni variabile: dimensione campione, media, deviazione standard ed errore standard della media. Per le differenze fra le medie: media, errore standard e intervallo di confidenza (è possibile specificare il livello di confidenza). Test: test di Levene di uguaglianza delle varianze e test *t* di uguaglianza delle medie per la varianza comune e la varianza separata.

Dati. I valori della variabile quantitativa desiderata si trovano in una singola colonna del file di dati. Viene utilizzata una variabile di raggruppamento che include due valori per suddividere i casi in due gruppi. La variabile di raggruppamento può essere numerica (valori quali 1 e 2, o 6,25 e 12,5) oppure una stringa breve (ad esempio *sì* e *no*). In alternativa, è possibile utilizzare una variabile quantitativa, ad esempio *età*, per suddividere i casi in due gruppi specificando un punto di divisione (il punto di divisione 21 suddivide la variabile *età* in un gruppo con meno di 21 anni e in un gruppo con più di 21 anni).

Assunzioni. Per il test *t* di uguaglianza della varianza, le osservazioni dovrebbero essere rappresentate da campioni indipendenti e casuali derivati da distribuzioni normali con la stessa varianza di popolazione. Per il test *t* di inuguaglianza della varianza, le osservazioni dovrebbero essere campioni indipendenti e casuali derivati da distribuzioni normali. Il test *t* per due campioni è sufficientemente robusto per le deviazioni dalla normalità. Durante la verifica grafica delle distribuzioni, controllare che siano simmetriche e che non siano presenti valori anomali.

Per ottenere un test T per campioni indipendenti

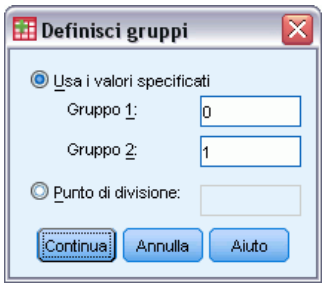
- Dai menu, scegliere:
  Analizza > Confronta medie > Test T: campioni indipendenti...

Figura 9-1
Finestra di dialogo Test T per campioni indipendenti

- Selezionare una o più variabili quantitative oggetto del test. Per ciascuna variabile viene calcolato un test *t* distinto.
- Selezionare una variabile di raggruppamento singola e fare clic su Definisci gruppi per specificare due codici per i gruppi che si desidera confrontare.
- Se necessario, fare clic su Opzioni per verificare in quale modo vengono considerati i dati mancanti e il livello dell'intervallo di confidenza.

Test T per campioni indipendenti: Definisci gruppi

Figura 9-2
Finestra di dialogo Definisci gruppi per le variabili numeriche



Per le variabili di raggruppamento numeriche, definire i due gruppi per il test *t* specificando due valori o un punto di divisione:

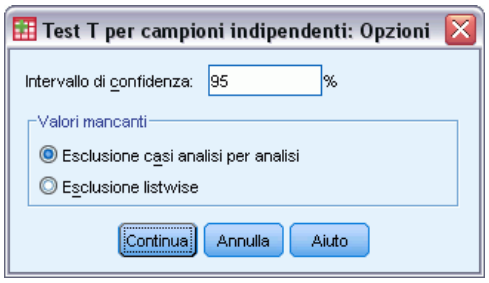
- **Usa i valori specificati.** Inserire un valore per Gruppo 1 e un altro valore per Gruppo 2. I casi con qualsiasi altro valore verranno esclusi dall'analisi. Non è necessario specificare numeri interi (ad esempio, 6,25 e 12,5 sono validi).
- **Punto di divisione.** Immettere un numero che suddivide i valori della variabile di raggruppamento in due insiemi Assegna i casi con valori minori o uguali a quello specificato a un gruppo e i casi con valori maggiori all'altro.

Figura 9-3
Finestra di dialogo Definisci gruppi per le variabili stringa

Per le variabili stringa di raggruppamento, inserire una stringa per Gruppo 1 e un'altra stringa per Gruppo 2, ad esempio *sì* e *no*. I casi che includono altre stringhe verranno esclusi dall'analisi.

Test T per campioni indipendenti: Opzioni

Figura 9-4
Finestra di dialogo Test T per campioni indipendenti: Opzioni



Intervallo di confidenza. Per impostazione predefinita, viene visualizzato un intervallo di confidenza del 95% per la differenza fra le medie. Immettere un valore compreso fra 1 e 99 per richiedere un livello di confidenza differente.

Valori mancanti. Se durante un test su più variabili si riscontra in alcune di esse la presenza di dati mancanti, è possibile indicare alla procedura i casi da includere (o da escludere):

- **Esclusione casi analisi per analisi.** Per ciascun test *t* vengono utilizzati tutti i casi con dati validi per le variabili verificate. Le dimensioni del campione possono variare in base al test.
- **Esclusione listwise.** Per ciascun test *t* vengono utilizzati solo i casi con dati validi per tutte le variabili prese in considerazione nei test *t*. La dimensione del campione è costante nei vari test.

T per campioni appaiati

La procedura Test T per campioni appaiati consente di confrontare le medie di due variabili per un singolo gruppo. La procedura calcola le differenze tra i valori delle due variabili per ciascun caso e viene verificato se la media è diversa da 0.

Esempio. In uno studio su pazienti con valori elevati della pressione sanguigna, a tutti i pazienti è stata misurata la pressione all'inizio dello studio, è stato somministrato un trattamento e quindi la misurazione è stata ripetuta. Per ciascun soggetto sono quindi disponibili due misurazioni, in genere denominate *precedente* e *successiva*. Questo test viene utilizzato anche per disegni relativi a studi di confronti tra coppie o di casi di controllo, in cui ciascun record del file di dati contiene la risposta per il paziente e quella del soggetto di controllo corrispondente. In uno studio sulla pressione sanguigna, è necessario che l'età dei pazienti trattati corrisponda a quella dei controlli (a un paziente di 75 anni deve corrispondere un membro del gruppo di controllo di 75 anni).

Statistiche. Per ogni variabile: media, dimensione campione, deviazione standard ed errore standard della media. Per ciascuna coppia di variabili: correlazione, differenza media tra le medie, test *t* e intervallo di confidenza per la differenza tra medie (è possibile specificare il livello di confidenza). Deviazione standard ed errore standard della differenza fra medie.

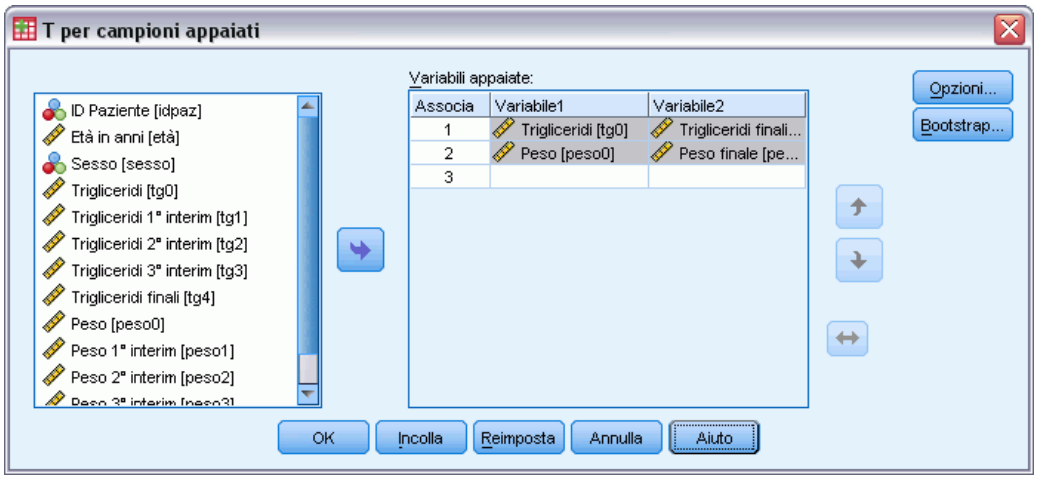
Dati. Per ciascun test appaiato, specificare due variabili quantitative (livello di misura in base a intervallo o a rapporto). In uno studio di confronti tra coppie o di casi di controllo, la risposta per ciascun soggetto del test e per il soggetto di controllo corrispondente deve trovarsi nello stesso caso all'interno del file di dati.

Assunzioni. Le osservazioni per ciascuna coppia devono essere effettuate nelle medesime condizioni. Le differenze tra medie devono essere distribuite normalmente. Le varianze di ciascuna variabile possono essere uguali o non uguali.

Per ottenere un test T per campioni appaiati

▶ Dai menu, scegliere:
Analizza > Confronta medie > Test T: campioni appaiati...

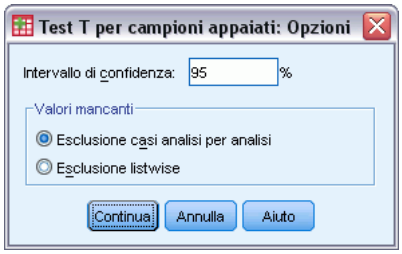
Figura 9-5
Finestra di dialogo Test T per campioni appaiati



- Selezionare una o più coppie di variabili
- Se necessario, fare clic su Opzioni per verificare in quale modo vengono considerati i dati mancanti e il livello dell'intervallo di confidenza.

Test T per campioni appaiati: Opzioni

Figura 9-6
Finestra di dialogo Test T per campioni appaiati: Opzioni



Intervallo di confidenza. Per impostazione predefinita, viene visualizzato un intervallo di confidenza del 95% per la differenza fra le medie. Immettere un valore compreso fra 1 e 99 per richiedere un livello di confidenza differente.

Valori mancanti. Se durante un test su più variabili si riscontra in alcune di esse la presenza di dati mancanti, è possibile indicare alla procedura i casi da includere (o da escludere):

- **Esclusione casi analisi per analisi.** Per ciascun test *t* vengono utilizzati tutti i casi con dati validi per la coppia di variabili verificate. Le dimensioni del campione possono variare in base al test.
- **Esclusione listwise.** Per ciascun test *t* vengono utilizzati solo i casi che includono dati validi per tutte le coppie di variabili verificate. La dimensione del campione è costante nei vari test.

Test T per un campione

La procedura Test T per un campione consente di verificare se la media di una singola variabile è diversa da una costante specificata.

Esempi. Un ricercatore può verificare se il quoziente d'intelligenza (QI) medio di un gruppo di studenti è diverso da 100. Oppure, un produttore di cereali può prelevare un campione di scatole dalla linea di produzione e verificare se il peso medio dei campioni è diverso da 0,7 kg con un livello di confidenza al 95%.

Statistiche. Per ciascuna variabile oggetto del test: media, deviazione standard ed errore standard della media. La differenza media fra ciascun valore e il valore oggetto del test ipotizzato, un test *t* che verifica che la differenza sia uguale a 0 e un intervallo di confidenza per la differenza (è possibile specificare il livello di confidenza).

Dati. Per confrontare i valori di una variabile quantitativa con un valore oggetto del test ipotizzato, scegliere una variabile quantitativa e immettere un valore oggetto del test ipotizzato.

Assunzioni. In questo test si presume che i dati siano distribuiti in modo normale. Il test è tuttavia sufficientemente robusto per le deviazioni dalla normalità.

Per ottenere un test T per un campione

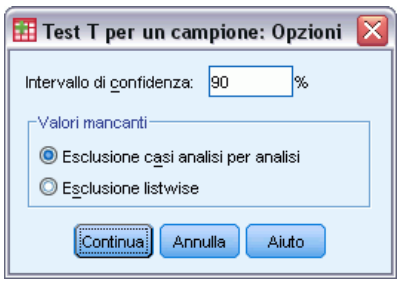
- Dai menu, scegliere:
  Analizza > Confronta medie > Test T: campione unico...

Figura 9-7
Finestra di dialogo Test T per un campione

- Selezionare una o più variabili da confrontare con lo stesso valore ipotizzato.
- Immettere un valore oggetto del test numerico rispetto al quale viene confrontata ciascuna media del campione.
- Se necessario, fare clic su Opzioni per verificare in quale modo vengono considerati i dati mancanti e il livello dell'intervallo di confidenza.

Test T per un campione: Opzioni

Figura 9-8
Finestra di dialogo Test T per un campione: Opzioni



Intervallo di confidenza. Per impostazione predefinita, viene visualizzato un intervallo di confidenza del 95% per la differenza fra la media e il valore oggetto del test ipotizzato. Immettere un valore compreso fra 1 e 99 per richiedere un livello di confidenza differente.

Valori mancanti. Se durante un test su più variabili si riscontra in alcune di esse la presenza di dati mancanti, è possibile indicare alla procedura i casi da includere (o da escludere):

- **Esclusione casi analisi per analisi.** Per ciascun test *t* vengono utilizzati tutti i casi con dati validi per le variabili verificate. Le dimensioni del campione possono variare in base al test.
- **Esclusione listwise.** Per ciascun tes*t* vengono utilizzati solo casi con dati validi per tutte le variabili prese in considerazione nei *t* test richiesti. La dimensione del campione è costante nei vari test.

Opzioni aggiuntive del comando T-TEST

Il linguaggio della sintassi dei comandi consente inoltre di:

- Effettuare test di T per un campione e per campioni indipendenti tramite un unico comando.
- Confrontare ciascuna variabile con le variabili dell'elenco in test accoppiati (con il sottocomando `PAIRS`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
10

ANOVA univariata

La procedura ANOVA univariata produce un'analisi della varianza univariata per una variabile dipendente quantitativa in base a una singola variabile fattore (indipendente). L'analisi della varianza consente di verificare l'ipotesi di uguaglianza di più medie. Questa tecnica è un'estensione del test *t* per due campioni.

Oltre a determinare le differenze tra le medie, è possibile individuare la media che differisce dalle altre. Esistono due tipi di test per il confronto tra le medie: contrasti a priori e test post hoc. I contrasti sono test impostati *prima* di eseguire l'esperimento, mentre i test post hoc vengono effettuati *dopo* l'esecuzione dell'esperimento. È inoltre possibile verificare i trend presenti tra le categorie.

Esempio. Le ciambelle assorbono quantità variabili di grassi a seconda della modalità di cottura. È stato condotto un esperimento che prevede l'utilizzo di tre tipi di grassi: olio di semi di arachide, olio di mais e strutto. L'olio di semi di arachide e l'olio di mais sono grassi insaturi, mentre lo strutto è un grasso saturo. Oltre a determinare se la quantità di grassi assorbita dipende dal tipo di grasso utilizzato, è possibile impostare un contrasto a priori per determinare se la quantità di grassi assorbita differisce per i grassi saturi e insaturi.

Statistiche. Per ciascun gruppo: numero di casi, media, deviazione standard, errore standard della media, valore minimo e massimo e intervallo di confidenza al 95% per la media. Test di omogeneità della varianza di Levene, tabella di analisi della varianza e test robusti dell'uguaglianza delle medie per ciascuna variabile dipendente, contrasti a priori definiti dall'utente e test di intervallo post hoc e confronti multipli: Bonferroni, Sidak, differenze significative di Tukey, GT2 di Hochberg, Gabriel, Dunnett, procedura di Ryan-Einot-Gabriel-Welsch basata su test *F*, (R-E-G-W *F*), test a intervalli di Ryan-Einot-Gabriel-Welsch (R-E-G-W *Q*), T2 di Tamhane, T3 di Dunnett, Games-Howell, *C* di Dunnett, test a intervalli multipli di Duncan, Student-Newman-Keuls (S-N-K), *b*di Tukey, Waller-Duncan, Scheffée differenza meno significativa.

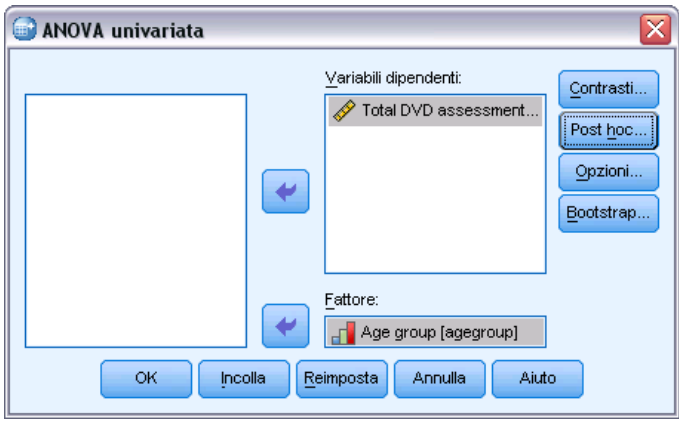
Dati. I valori delle variabili fattore devono essere interi e la variabile dipendente deve essere quantitativa (livello di misura per intervallo).

Assunzioni. Ciascun gruppo è un campione casuale indipendente prelevato da una popolazione normale. L'analisi della varianza è uno stimatore robusto degli scostamenti dalla normalità, anche se i dati devono essere simmetrici. I gruppi devono provenire da popolazioni con varianze uguali. Per verificare questa ipotesi, utilizzare il test dell'omogeneità della varianza di Levene.

Per ottenere un'analisi della varianza univariata

- Dai menu, scegliere:
  Analizza > Confronta medie > ANOVA univariata...

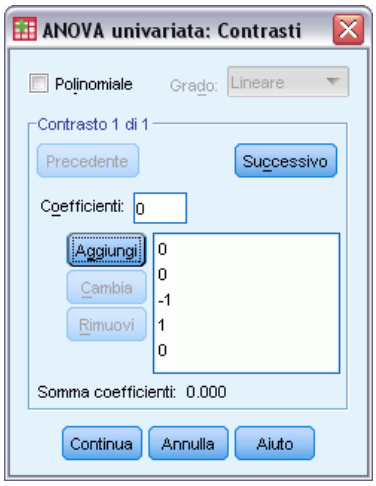
Figura 10-1
Finestra di dialogo ANOVA univariata



- Selezionare una o più variabili dipendenti.
- Selezionare una singola variabile fattore indipendente.

ANOVA univariata: Contrasti

Figura 10-2
Finestra di dialogo ANOVA univariata: Contrasti



È possibile suddividere le somme dei quadrati fra gruppi in componenti di trend oppure specificare contrasti a priori.

Polinomiale. Consente di suddividere le somme dei quadrati tra gruppi in componenti di trend. È possibile verificare un trend della variabile dipendente in tutti i livelli ordinati della variabile fattore. Ad esempio, è possibile verificare un trend lineare (crescente o decrescente) nei salari in tutti i livelli ordinati del grado di salario più elevato.

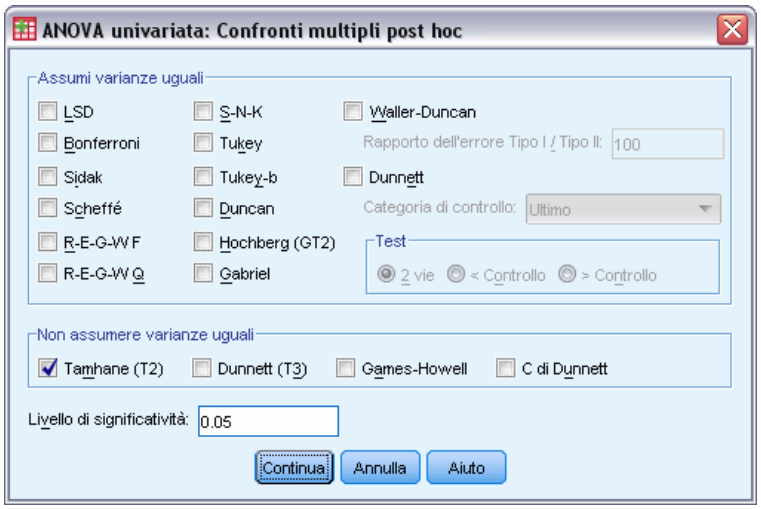
- **Grado.** È possibile scegliere un termine polinomiale di ordine 1, 2, 3, 4 o 5.

Coefficienti. Contrasti a priori definiti dall'utente da verificare con la statistica *t*. Specificare un coefficiente per ciascun gruppo (categoria) della variabile fattore e quindi fare clic su Aggiungi dopo aver inserito ciascuna voce. I nuovi valori verranno aggiunti alla fine dell'elenco dei coefficienti. Per specificare altri insiemi di contrasti, fare clic su Successivo. Utilizzare Successivo e Precedente per spostarsi tra gli insiemi di contrasti.

L'ordine dei coefficienti è importante in quanto corrisponde all'ordine crescente dei valori delle categorie della variabile fattore. Il primo coefficiente dell'elenco corrisponde al valore di gruppo minimo della variabile fattore e l'ultimo coefficiente corrisponde al valore massimo. Ad esempio, se sono presenti sei categorie della variabile fattore, i coefficienti –1, 0, 0, 0, 0,5 e 0,5 contrastano il primo gruppo con il quinto e il sesto gruppo. Per la maggior parte delle applicazioni la somma dei coefficienti deve essere 0. È possibile utilizzare anche insiemi la cui somma è diversa da 0, ma verrà visualizzato un messaggio di avvertimento.

ANOVA univariata: Test Post Hoc

Figura 10-3
Finestra di dialogo ANOVA univariata: Confronti multipli Post Hoc



Dopo aver determinato l'esistenza di differenze tra le medie, i test post hoc di intervalli e i confronti a coppie multipli consentono di determinare quale media differisce dalle altre. I test a intervalli multipli consentono di identificare sottoinsiemi omogenei di medie che non differiscono le une dalle altre. Grazie ai confronti a coppie multipli è possibile verificare la differenza tra ciascuna coppia di medie e ottenere una matrice in cui gli asterischi indicano le medie di gruppo con differenze significative e un livello alfa 0,05.

Assumi varianze uguali

Il test delle differenze significative di Tukey, il GT2 di Hochberg, il test di Gabriel e il test di Scheffé sono test a confronti e intervalli multipli. Sono disponibili altri test di intervalli, ovvero *b* di Tukey, S-N-K (Student-Newman-Keuls), Duncan, R-E-G-W *F* (test *F* di Ryan-Einot-Gabriel-Welsch), R-E-G-W *Q* (test a intervalli di Ryan-Einot-Gabriel-Welsch) e Waller-Duncan. I test a confronti multipli disponibili sono i seguenti: Bonferroni, test delle

differenze significative di Tukey, Sidak, Gabriel, Hochberg, Dunnett, Scheffé e LSD (differenza meno significativa).

- **LSD.** Usa i test t per eseguire tutti i confronti a coppie tra medie di gruppo. Non viene apportata alcuna correzione al tasso di errore per i confronti multipli.
- **Bonferroni.** Usa i test t per eseguire confronti a coppie tra medie di gruppo, ma controlla il tasso di errore globale impostando il tasso di errore di ogni test al tasso di errore sperimentale diviso per il numero totale dei test. In questo modo il livello di significatività osservato è corretto tenendo conto che si stanno effettuando confronti multipli.
- **Sidak.** Test per confronti a coppie multipli basato sul test t. Effettua la correzione del livello di significatività per confronti multipli e fornisce una banda più stretta rispetto al test di Bonferroni.
- **Scheffé.** Effettua confronti congiunti a coppie simultanei per tutte le possibili coppie di medie. Utilizza la distribuzione di campionamento F. Può essere usato per esaminare tutte le possibili combinazioni lineari di medie di gruppo, non solo i confronti a coppie.
- **R-E-G-W F.** La procedura di Ryan-Einot-Gabriel-Welsch, basata su un test F.
- **R-E-G-W Q.** La procedura di Ryan-Einot-Gabriel-Welsch, basata su un intervallo studentizzato.
- **S-N-K.** Effettua tutti i confronti a coppie fra medie usando la distribuzione di intervallo studentizzata. Per dimensioni campionarie uguali, confronta anche coppie di medie entro sottoinsiemi omogenei, utilizzando una procedura pairwise. Le medie vengono ordinate dalla più alta alla più bassa e vengono verificate per prime le differenze estreme.
- **Tukey.** Usa la statistica di intervallo studentizzato per effettuare tutti i confronti a coppie tra gruppi. Imposta il tasso di errore sperimentale al valore del tasso di errore per l'insieme di tutti i confronti per coppie.
- **Tukey-b.** Utilizza la distribuzione di intervallo studentizzato per effettuare confronti a coppie tra gruppi. Il valore critico è la media fra il corrispondente valore per il test HSD di Tukey e quello del test di Student-Newman-Keuls.
- **Duncan.** Effettua confronti a coppie usando lo stesso ordine di confronti per passi usato nel test di Student Newman Keuls, impostando un livello di soglia per il tasso di errore valido per l'insieme dei test, invece di un tasso di errore diverso per ciascun test. Usa la statistica dell'intervallo studentizzato.
- **Hochberg (GT2).** Test per confronti multipli o per intervallo basato sul modulo studentizzato. Simile al test HSD di Tukey.
- **Gabriel.** Test per confronti a coppie basato sul modulo studentizzato, generalmente più indicato del test GT2 quando le celle hanno dimensioni diverse. Se la variabilità delle dimensioni delle celle risulta molto alta, il test di Gabriel può diventare poco conservativo.
- **Waller-Duncan.** Test per confronti multipli basato su una statistica t. Utilizza un approccio bayesiano.
- **Dunnett.** Test per confronti a coppie multipli basato sul test t, che confronta un insieme di trattamenti con un'unica media di controllo. L'ultima categoria è la categoria di controllo predefinita. In alternativa, è possibile scegliere la prima categoria. I test a due sensi consentono di verificare che la media in qualsiasi livello del fattore (ad eccezione della categoria di controllo) non sia uguale a quella della categoria di controllo. I test di <controllo consentono di verificare se la media di qualsiasi livello del fattore sia minore di quella della

categoria di controllo. I test di >controllo consentono di verificare se la media di qualsiasi livello del fattore sia maggiore di quella della categoria di controllo.

Non assumere varianze uguali

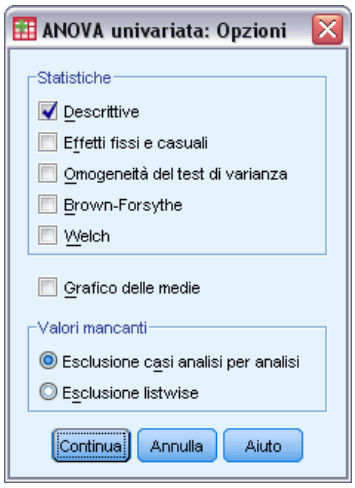
Sono disponibili test a confronti multipli che non ipotizzano varianze uguali, ovvero T2 di Tamhane, T3 di Dunnett, Games-Howell e *C* di Dunnett.

- **Tamhane (T2).** Test per confronti a coppie basato sul test t, appropriato quando le varianze non sono uguali.
- **Dunnett (T3).** Test per confronti a coppie basato sul modulo studentizzato. appropriato quando le varianze non sono uguali.
- **Games-Howell.** Test per confronti a coppie non molto conservativo, appropriato quando le varianze non sono uguali.
- **C di Dunnett.** Test per confronti a coppie basato sull'intervallo studentizzato, appropriato quando le varianze non sono uguali.

Nota: Per interpretare più rapidamente l'output dei test post hoc è consigliabile deselezionare Nascondi righe e colonne vuote nella finestra di dialogo Proprietà tabella (in una tabella pivot attivata, scegliere Proprietà tabella dal menu Formato).

ANOVA univariata: Opzioni

Figura 10-4
Finestra di dialogo ANOVA univariata: Opzioni



Statistiche. Consente di scegliere una o più delle seguenti opzioni:

- **Descrittive.** Consente di calcolare il numero di casi, la media, la deviazione standard, l'errore standard della media, il valore minimo e massimo e gli intervalli di confidenza al 95% per ciascuna variabile dipendente di ciascun gruppo.

- **Effetti fissi e casuali.** Consente di visualizzare la deviazione standard, l'errore standard e l'intervallo di confidenza del 95% per il modello degli effetti fissi e l'errore standard, l'intervallo di confidenza del 95% e la stima della varianza tra componenti per il modello degli effetti casuali.
- **Omogeneità del test di varianza.** Consente di calcolare il test di Levene per verificare l'uguaglianza tra gruppi di variabili. Questo test non si basa sull'ipotesi di normalità.
- **Brown-Forsythe.** Consente di calcolare la statistica di Brown-Forsythe per verificare l'uguaglianza delle medie dei gruppi. Questa statistica è preferibile alla statistica *F* nel caso in cui non sia valida l'ipotesi di uguaglianza della varianza.
- **Welch.** Consente di calcolare la statistica di Welch per verificare l'uguaglianza delle medie dei gruppi. Questa statistica è preferibile alla statistica *F* nel caso in cui non sia valida l'ipotesi di uguaglianza della varianza.

Grafico delle medie. Consente di visualizzare un grafico che rappresenta le medie dei sottogruppi (le medie di ciascun gruppo definite dai valori della variabile fattore).

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi analisi per analisi.** I casi con valori mancanti per la variabile dipendente o fattore per una particolare analisi non verranno utilizzati in tale analisi. Non verranno utilizzati nemmeno i casi che non rientrano nell'intervallo specificato per la variabile fattore.
- **Esclusione listwise.** I casi con valori mancanti per la variabile fattore o qualsiasi variabile dipendente inclusa nella lista delle variabili dipendenti nella finestra di dialogo principale verranno esclusi da tutte le analisi. Se non sono state specificate più variabili dipendenti, l'opzione non produrrà alcun effetto.

Opzioni aggiuntive del comando ONEWAY

Il linguaggio della sintassi dei comandi consente inoltre di:

- Ottenere statistiche con effetti fissi e casuali. Deviazione standard, errore standard della media e intervalli di confidenza al 95% per il modello con effetti fissi. Errore standard, intervalli di confidenza al 95% e stima della varianza dei componenti per il modello con effetti casuali (con il sottocomando `STATISTICS=EFFECTS`).
- Specificare i livelli alfa per la differenza meno significativa nei test per confronti a coppie multipli di Bonferroni, Duncan e Scheffé (con il sottocomando `RANGES`).
- Scrivere una matrice di medie, deviazioni standard e frequenze, oppure leggere una matrice di medie, frequenze, varianze raggruppate e gradi di libertà per le varianze raggruppate. Queste matrici possono essere usate al posto dei dati grezzi per effettuare un'analisi univariata della varianza (con il sottocomando `MATRIX`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
11

Analisi GLM univariato

La procedura GLM univariato consente di eseguire un'analisi di regressione e un'analisi della varianza per una variabile dipendente tramite uno o più fattori e/o variabili. Le variabili fattore suddividono la popolazione in gruppi. Con questa procedura GLM (modello lineare generalizzato, General Linear Model) è possibile verificare ipotesi nulle relative agli effetti di altre variabili sulle medie di vari raggruppamenti di una sola variabile dipendente. È possibile analizzare le interazioni tra fattori e gli effetti di singoli fattori, alcuni dei quali possono essere casuali. È inoltre possibile includere gli effetti delle covariate e le interazioni tra covariate e fattori. Nell'analisi di regressione, le variabili indipendenti (stimatori) vengono specificate come covariate.

È possibile verificare sia modelli bilanciati che modelli non bilanciati. Un disegno è bilanciato se ciascuna cella del modello include lo stesso numero di casi. Oltre alla verifica delle ipotesi, la procedura GLM univariato consente di ottenere stime dei parametri.

Per la verifica di ipotesi sono disponibili contrasti a priori usati di frequente. Dopo che da un test *F* globale è risultata una certa significatività, è inoltre possibile eseguire test post hoc per valutare le differenze tra medie specifiche. L'opzione Medie marginali stimate consente di ottenere stime dei valori medi previsti delle celle incluse nel modello. I grafici di profilo, o grafici di interazione, di tali medie consentono di visualizzare in modo semplice alcune delle relazioni.

Residui, valori attesi, distanza di Cook e valori d'influenza possono essere salvati come variabili nel file di dati per la verifica di ipotesi.

Minimi quadrati ponderati consente di specificare una variabile per l'assegnazione di pesi diversi alle osservazioni per un'analisi di minimi quadrati ponderati (WLS), in alcuni casi per compensare la diversa precisione della misura.

Esempio. Per diversi anni vengono raccolti i dati relativi ai singoli partecipanti alla maratona di Chicago. Il tempo impiegato da ciascun partecipante per completare la maratona è la variabile dipendente. Altri fattori presi in considerazione sono le condizioni metereologiche (freddo, caldo o temperatura moderata), il numero di mesi di allenamento, il numero di maratone corse in precedenza e il sesso. L'età è considerata una covariata. Dallo studio può risultare che il sesso rappresenta un effetto significativo, così come l'interazione tra sesso e condizioni meteorologiche.

Metodi. Per la valutazione di ipotesi diverse è possibile usare la somma dei quadrati Tipo I, Tipo II, Tipo III e Tipo IV. Il metodo predefinito è il Tipo III.

Statistiche. I seguenti test post hoc di intervalli e confronti multipli: Differenza meno significativa (LSD), Bonferroni, Sidak, Scheffé, Ryan-Einot-Gabriel-Welsch multiplo basato su test *F*, Ryan-Einot-Gabriel-Welsch a intervallo multiplo, Student-Newman-Keuls, differenze significative di Tukey, *b* di Tukey, Duncan, Hochberg (GT2), Gabriel, *t* di Waller-Duncan, Dunnett (a una e a due vie), Tamhane (T2), Dunnett (T3), Games-Howell e *C* di Dunnett. Statistiche descrittive: medie osservate, deviazioni standard e conteggi per tutte le variabili dipendenti di tutte le celle. Test di Levene di omogeneità della varianza.

Grafici. Grafici di variabilità vs. densità, dei residui e di profilo (interazione).

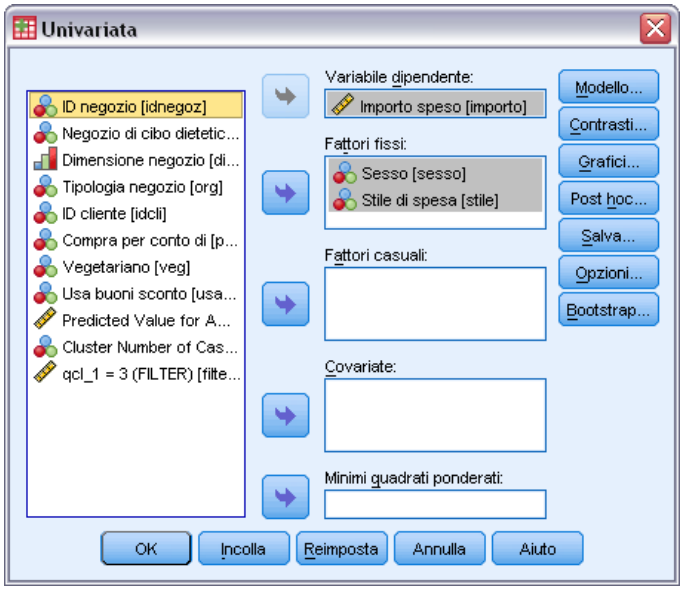
Dati. La variabile dipendente è quantitativa. I fattori sono categoriali. Vi possono essere associati valori numerici o valori stringa composti da un massimo di otto caratteri. Le covariate sono variabili quantitative correlate alla variabile dipendente.

Assunzioni. I dati sono costituiti da un campione casuale derivato da una popolazione normale in cui tutte le varianze di cella sono uguali. L'analisi della varianza è uno stimatore robusto degli scostamenti dalla normalità, anche se i dati devono essere simmetrici. Per la verifica di ipotesi, è possibile usare test di omogeneità della varianza e i grafici di variabilità vs. intensità. È inoltre possibile esaminare residui e grafici dei residui.

Per ottenere tabelle di GLM univariato

- Dai menu, scegliere:
  Analizza > Modello lineare generalizzato > Univariata...

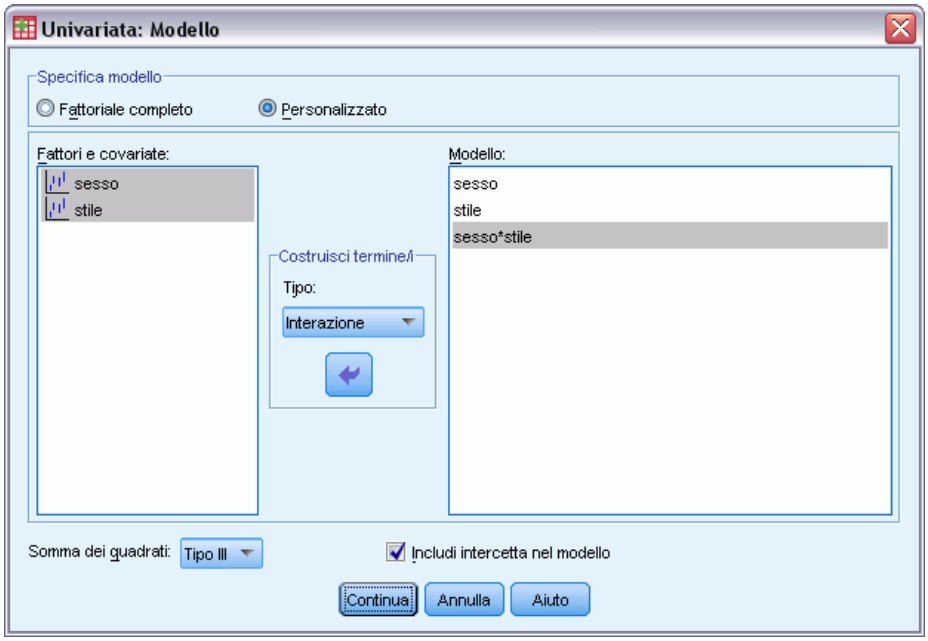
Figura 11-1
Finestra di dialogo GLM univariato



- Selezionare una variabile dipendente.
- Selezionare le variabili per l'opzione Fattori fissi, Fattori casuali o Covariate, a seconda dei dati in uso.
- È inoltre possibile utilizzare l'opzione Minimi quadrati ponderati per specificare l'analisi dei minimi quadrati ponderati. Casi con valore 0, negativo o mancante per la variabile di ponderazione saranno esclusi dall'analisi. Non è possibile utilizzare come variabile di ponderazione una variabile già inclusa nel modello.

GLM – Univariato: Modello

Figura 11-2
Finestra di dialogo Univariata: Modello



Specifica modello. Un modello fattoriale completo contiene tutti gli effetti principali dei fattori e delle covariate e tutte le interazioni fattore per fattore. Non contiene interazioni di covariate Selezionare Personalizzato per specificare un solo sottoinsieme di interazioni o interazioni tra fattori e covariate. È necessario indicare tutti i termini da includere nel modello.

Fattori e covariate. I fattori e le covariate sono elencati.

Modello. Il modello varia in base alla natura dei dati in uso. Dopo aver selezionato Personalizzato, è possibile selezionare gli effetti principali e le interazioni desiderate per l'analisi da eseguire.

Somma dei quadrati. Metodo per il calcolo della somma dei quadrati. Il metodo Somma dei quadrati in genere utilizzato con modelli bilanciati o non bilanciati privi di celle mancanti è il Tipo III.

Includi l'intercetta nel modello. L'intercetta viene in genere inclusa nel modello. Se è possibile presumere che i dati passino attraverso l'origine, l'intercetta può essere esclusa.

Costruisci termini

Per i fattori e le covariate selezionati:

Interazione. Consente di creare il termine di interazione di livello maggiore rispetto a tutte le variabili selezionate. È l'impostazione predefinita.

Effetti principali. Consente di creare un termine di effetti principali per ciascuna variabile selezionata.

Tutti 2-vie. Consente di creare tutte le possibili interazioni a due vie delle variabili selezionate.

Tutti 3-vie. Consente di creare tutte le possibili interazioni a tre vie delle variabili selezionate.

Tutti 4-vie. Consente di creare tutte le possibili interazioni a quattro vie delle variabili selezionate.

Tutti 5-vie. Consente di creare tutte le possibili interazioni a cinque vie delle variabili selezionate.

Somma dei quadrati

Per il modello è possibile scegliere un tipo di somma dei quadrati. Il Tipo III, il tipo predefinito, è quello usato più di frequente.

Tipo I. Questo metodo è definito anche scomposizione gerarchica del metodo Somma dei quadrati. Ciascun termine viene corretto solo per i termini del modello che lo precedono. Il metodo Somma dei quadrati Tipo I è in genere usato con i seguenti elementi:

- Un modello ANOVA bilanciato in cui gli effetti principali vengono specificati prima degli effetti di interazione di ordine 1, ciascuno dei quali viene a sua volta specificato prima degli effetti di interazione di ordine 2 e così via.
- Un modello di regressione polinomiale in cui qualsiasi termine di ordine più basso è specificato prima dei termini di ordine più elevato.
- Un modello nidificato in modo puro in cui il primo effetto specificato è nidificato nel secondo, il quale è a sua volta nidificato nel terzo e così via. Questo tipo di nidificazione può essere specificato esclusivamente tramite la sintassi.

Tipo II. Questo metodo consente di calcolare le somme dei quadrati di un effetto del modello corretto per tutti gli altri effetti "appropriati". È considerato appropriato un effetto corrispondente a tutti gli effetti che non includono l'effetto in esame. Il metodo Somma dei quadrati Tipo II è in genere usato con i seguenti elementi:

- Un modello ANOVA bilanciato.
- Qualsiasi modello che include solo effetti principali del fattore.
- Qualsiasi modello di regressione.
- Un disegno nidificato in modo puro. Questo tipo di nidificazione può essere specificato tramite la sintassi.

Tipo III. Tipo predefinito. Questo metodo consente di calcolare le somma dei quadrati di un effetto del disegno come la somma dei quadrati corretta per qualsiasi altro effetto che non lo include e ortogonale rispetto agli eventuali effetti che lo contengono. Il vantaggio associato a questo tipo di somme dei quadrati è che non varia al variare delle frequenze di cella, a condizione che la forma generale di stimabilità rimanga costante. È pertanto considerato utile per modelli non bilanciati privi di celle mancanti. In un disegno fattoriale privo di celle mancanti, questo metodo equivale alla tecnica dei quadrati delle medie ponderate di Yates. Il metodo Somma dei quadrati Tipo III è in genere usato con i seguenti elementi:

- I modelli elencati per il Tipo I e il Tipo II.
- Qualsiasi modello bilanciato o non bilanciato e privo di celle vuote.

Tipo IV. Questo metodo è specifico per situazioni con celle mancanti. Per qualsiasi effetto *F* del disegno, se *F* non è incluso in nessun altro effetto, allora Tipo IV = Tipo III = Tipo II. Se invece *F* è incluso in altri effetti, con il Tipo IV i contrasti creati tra i parametri in *F* vengono distribuiti

equamente tra tutti gli effetti di livello superiore. Il metodo Somma dei quadrati Tipo IV viene in genere usato con i seguenti elementi:

- I modelli elencati per il Tipo I e il Tipo II.
- Qualsiasi modello bilanciato e non bilanciato contenente celle vuote.

GLM – Univariato: Contrasti

Figura 11-3
Finestra di dialogo Univariata: Contrasti

I contrasti consentono di verificare il grado di differenza tra i livelli di un fattore. È possibile specificare un contrasto per ciascun fattore del modello (in un modello a misure ripetute, un contrasto per ciascun fattore tra soggetti). I contrasti rappresentano combinazioni lineari dei parametri.

La verifica di ipotesi è basata sull'ipotesi nulla **LB** = 0, dove **L** è la matrice dei coefficienti di contrasto e **B** è il vettore dei parametri. Quando si specifica un contrasto, viene creata una matrice **L**. Le colonne della matrice **L** corrispondenti al fattore corrispondono al contrasto. Le altre colonne vengono corrette in modo che la matrice **L** possa essere stimata.

L'output include una statistica *F* per ciascun insieme di contrasti. Per le differenze dei contrasti vengono inoltre visualizzati gli intervalli di confidenza simultanei di tipo Bonferroni basati su una distribuzione *t* di Student.

Contrasti disponibili

Sono disponibili i contrasti deviazione, semplici, differenza, Helmert, ripetuti e polinomiali. Per i contrasti deviazione e i contrasti semplici, è possibile stabilire se la categoria di riferimento corrisponde alla prima o all'ultima categoria.

Tipi di contrasto

Deviazione. Consente di confrontare la media di ciascun livello, a eccezione di una categoria di riferimento, con la media di tutti i livelli (media principale). L'ordine dei livelli del fattore può essere un ordine qualsiasi.

Semplice. Consente di confrontare la media di ciascun livello con la media di un livello specifico. Questo tipo di contrasto risulta utile quando è disponibile un gruppo di controllo. Come categoria di riferimento, è possibile scegliere la prima o l'ultima categoria.

Differenza. Consente di confrontare la media di ciascun livello (a eccezione del primo) con la media dei livelli precedenti. Questo tipo di contrasto è a volte definito contrasto inverso di Helmert.

Helmert. Consente di confrontare la media di ciascun livello del fattore (a eccezione dell'ultimo) con la media dei livelli successivi.

Ripetuto. Consente di confrontare la media di ciascun livello (a eccezione dell'ultimo) con la media del livello successivo.

Polinomiale. Consente di confrontare l'effetto lineare, quadratico, cubico e così via. Tutte le categorie del primo grado di libertà includono l'effetto lineare, quelle del secondo includono l'effetto quadratico e così via. Questi contrasti sono spesso usati per la stima di trend polinomiali.

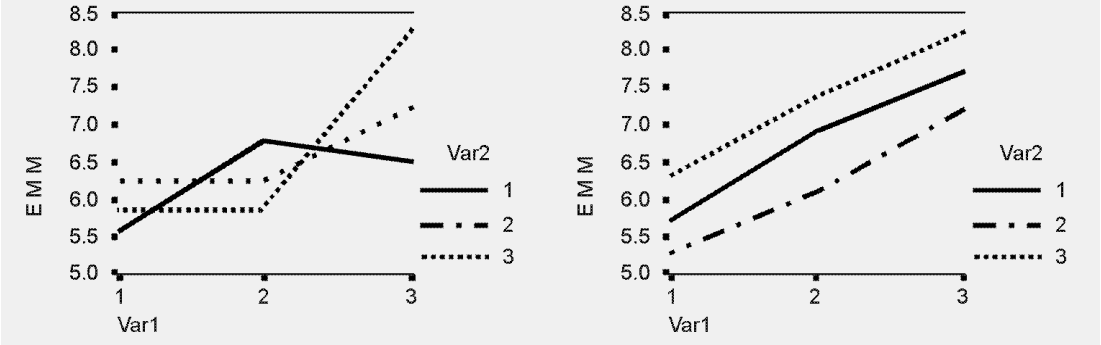
GLM - Univariato: Profili

Figura 11-4
Finestra di dialogo Univariata: Profili

I profili, o grafici di interazione, risultano utili per il confronto delle medie marginali di un modello. Un profilo è un grafico lineare in cui ciascun punto indica la media marginale stimata di una variabile dipendente (corretta per le covariate) in corrispondenza di un solo livello di un fattore. È possibile utilizzare i livelli di un secondo fattore per creare linee distinte. È possibile utilizzare ciascun livello di un terzo fattore per creare un grafico distinto. Tutti gli eventuali fattori casuali e fissi sono disponibili per i grafici. In analisi multivariate i grafici di profilo vengono creati per ciascuna variabile dipendente. Nei grafici di profilo per un'analisi a misure ripetute è possibile includere sia fattori tra soggetti che fattori entro soggetti. GLM Multivariato e GLM Misure ripetute sono disponibili solo se è stata installata l'opzione Advanced Statistics.

Il profilo di un fattore mostra se le medie marginali stimate aumentano o diminuiscono tra i vari livelli. Nel caso di due o più fattori, le linee parallele indicano che tra i fattori non esiste alcuna interazione, ovvero che è possibile analizzare i livelli di un solo fattore. Le linee che si incrociano indicano invece che esiste un'interazione.

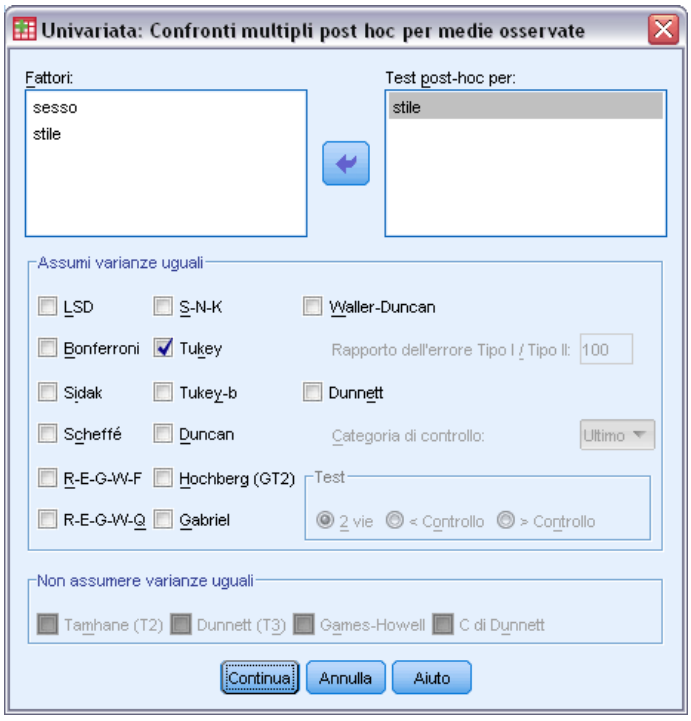
Figura 11-5
Grafico con linee non parallele (a sinistra) e grafico con linee parallele (a destra)



I grafici definiti tramite la selezione dei fattori per l'asse orizzontale e, se lo si desidera, dei fattori di linee e di grafici separati devono essere inclusi nell'elenco dei grafici.

GLM - Univariato: Confronti post hoc

Figura 11-6
Finestra di dialogo Post Hoc



Test per confronti multipli post hoc. Dopo aver determinato l'esistenza di differenze tra le medie, i test post hoc di intervalli e i confronti a coppie multipli consentono di determinare quale media differisce dalle altre. I confronti vengono eseguiti su valori a cui non è stata apportata alcuna correzione. Questi test vengono usati solo con fattori fissi tra soggetti. Nella procedura GLM a misure ripetute, questi test non sono disponibili se non sono presenti fattori tra soggetti e i test per confronti multipli post hoc vengono eseguiti per la media tra i livelli dei fattori entro soggetti.

Per la procedura GLM multivariato, i test post hoc vengono eseguiti separatamente per ciascuna variabile dipendente. GLM Multivariato e GLM Misure ripetute sono disponibili solo se è stata installata l'opzione Advanced Statistics.

I test per confronti multipli usati più di frequente sono il test di Bonferroni e il test delle differenze significative di Tukey. Il **test di Bonferroni**, basato sulla statistica *t* di Student, consente di correggere il livello di significatività osservato in base al fatto che vengono eseguiti confronti multipli. Il **test t di Sidak** corregge inoltre il livello di significatività ed è più restrittivo del test di Bonferroni. Il **test delle differenze significative di Tukey** utilizza la statistica di intervallo studentizzato per effettuare tutti i confronti a coppie tra gruppi e imposta il tasso di errore sperimentale sul valore del tasso di errore per l'insieme di tutti i confronti per coppie. Quando si eseguono test su un elevato numero di coppie di medie, il test delle differenze significative di Tukey risulta più efficace rispetto al test di Bonferroni. Nel caso di un numero limitato di coppie, risulta invece più efficace il test di Bonferroni.

Il test di **Hochberg (GT2)** è simile al test delle differenze significative di Tukey, ma utilizza il modulo massimo studentizzato. Il test di Tukey risulta in genere più efficace. Anche il **test dei confronti a coppie di Gabriel** utilizza il modulo massimo studentizzato ed è in genere più indicativo del test di Hochberg (GT2) quando le dimensioni delle celle sono diverse. Se la variabilità delle dimensioni delle celle risulta molto alta, il test di Gabriel può diventare poco conservativo.

Il **test t per confronti multipli a coppie di Dunnett** confronta un insieme di trattamenti con una media di controllo singola. L'ultima categoria è la categoria di controllo predefinita. In alternativa, è possibile scegliere la prima categoria. È inoltre possibile scegliere un test a 2 vie oppure a 1 via. Per verificare che la media in qualsiasi livello del fattore (a eccezione della categoria di controllo) non sia uguale a quella della categoria di controllo, è necessario usare un test a due sensi. Per verificare se la media di qualsiasi livello del fattore è minore di quella della categoria di controllo, selezionare < Controllo. In modo analogo, per verificare se la media di qualsiasi livello del fattore è maggiore di quella della categoria di controllo, selezionare > Control.

Ryan, Einot, Gabriel e Welsch (R-E-G-W) hanno sviluppato due test a intervalli decrescenti multipli. Le procedure a multipli decrescenti verificano in primo luogo se tutte le medie sono uguali. Se le medie non risultano tutte uguali, il test di uguaglianza viene eseguito su un sottoinsieme di medie. Il test **R-E-G-W F** è basato su un test *F*, mentre **R-E-G-W Q** è basato sull'intervallo studentizzato. Questi test risultano più efficaci rispetto ai test a intervallo multiplo di Duncan e Student-Newman-Keuls, che sono pure procedure a intervalli decrescenti multipli. È tuttavia consigliabile non usarli con celle di dimensioni non uguali.

Quando le varianze non sono uguali, è necessario utilizzare il test **Tamhane (T2)** (test per confronti a coppie conservativo basato su un test*t*), il test di **Dunnett T3** (test per confronti a coppie basato sul modulo studentizzato), il test per confronti a coppie di **Games-Howell** (a volte poco conservativo) o il test **C di Dunnett**(test per confronti a coppie basato sull'intervallo studentizzato). Notare che questi test non sono validi e non verranno eseguiti se il modello contiene più fattori.

Il test a intervallo multiplo di Duncan, il test di Student-Newman-Keuls (**S-N-K**) e il test **b di Tukey** sono test per intervallo che classificano le medie raggruppate e calcolano un valore di intervallo. Questi test sono usati meno frequentemente dei test descritti in precedenza.

Il **test t di Waller-Duncan** utilizza un approccio bayesiano. Si tratta di un test a intervallo che usa la media armonica della dimensione campione nel caso di dimensioni campione non uguali.

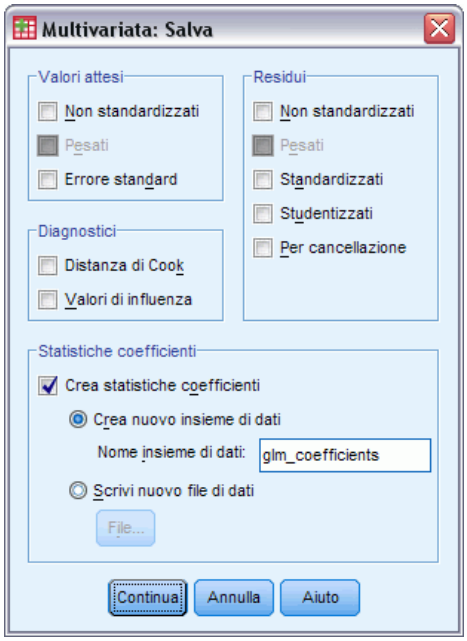
Il livello di significatività del test di **Scheffé** consente la verifica di tutte le possibili combinazioni lineari delle medie di gruppo e non dei soli confronti a coppie disponibili in questa funzione. Il test di Scheffé risulta pertanto più conservativo rispetto ad altri test, ovvero per ottenere un livello sufficiente di significatività, è richiesta una differenza maggiore tra le medie.

Il test per confronti a coppie multipli Differenza meno significativa o **LSD**, è equivalente a più test *t* tra tutte le coppie di gruppi. Lo svantaggio di questo test è che non viene eseguito alcun tentativo di correzione del livello di significatività osservata per confronti multipli.

Test visualizzati. I confronti a coppie sono disponibili per i test LSD, Sidak, Bonferroni, Games-Howell, Tamhane (T2) e (T3), test *C* di Dunnett e Dunnett (T3). Per i test S-N-K, *b* di Tukey, Duncan, R-E-G-W *F*, R-E-G-W *Q* e Waller sono disponibili sottoinsiemi omogenei per test per intervallo. Il test delle differenze significative di Tukey, i test Hochberg (GT2), Gabriel e Scheffé sono sia test per confronti multipli che test a intervallo.

GLM - Univariato: Salva

Figura 11-7
Salva



È possibile salvare i valori attesi dal modello, le misure correlate e i residui come nuove variabili nell'Editor dei dati. Molte di queste variabili possono essere usate per l'esame di ipotesi sui dati. Per salvare i valori in modo da poterli usare in un'altra sessione IBM® SPSS® Statistics, è necessario salvare il file di dati corrente.

Valori attesi. Valori attesi dal modello per ciascun caso.

- **Non standardizzati.** I valori risultanti dal modello per la variabile dipendente e per ciascun caso.

- **Pesati.** I valori attesi ponderati non standardizzati. Disponibile solo se è stata selezionata una variabile WLS.
- **Errore standard.** Una stima della deviazione standard del valore medio della variabile dipendente per i casi che hanno gli stessi valori delle variabili indipendenti.

Diagnostici. Misure per l'identificazione dei casi con combinazioni di valori insolite per le variabili indipendenti e dei casi che possono avere una notevole influenza sul modello.

- **Distanza di Cook.** Una misura di quanto cambierebbero i residui di tutti i casi se un particolare caso fosse escluso dal calcolo dei coefficienti di regressione. Valori alti indicano che l'esclusione di un caso dal calcolo dei coefficienti di regressione ne modificherebbe sostanzialmente il valore.
- **Valori di influenza.** Valori di influenza non centrati. Una misura dell'influenza di ciascun caso sull'adattamento di un modello di regressione.

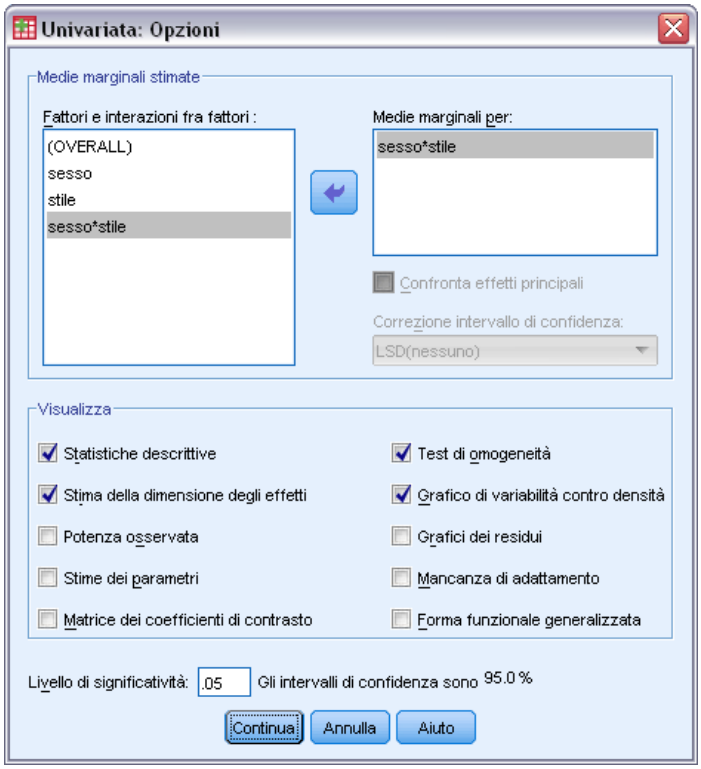
Residui. Un residuo non standardizzato corrisponde al valore effettivo della variabile dipendente diminuito del valore atteso dal modello. Sono inoltre disponibili residui standardizzati, studentizzati ed eliminati. Se è stata selezionata una variabile WLS, saranno inoltre disponibili residui non standardizzati pesati.

- **Non standardizzati.** La differenza tra un valore osservato e il valore stimato dal modello.
- **Pesati.** I residui ponderati non standardizzati. Disponibile solo se è stata selezionata una variabile WLS.
- **Standardizzati.** Il residuo diviso per una stima della deviazione standard. Il residuo standardizzato, conosciuto anche come residuo di Pearson, ha media 0 e deviazione standard 1.
- **Studentizzati.** Il residuo diviso per una stima della sua deviazione standard che varia da caso a caso, a seconda della distanza tra i valori assunti per questo caso dalle variabili indipendenti e le medie delle variabili indipendenti.
- **Per cancellazione.** Il residuo per un caso se quel caso venisse escluso dal calcolo dei coefficienti di regressione. È la differenza tra il valore della variabile dipendente e il valore stimato corretto.

Statistiche dei coefficienti. Scrive una matrice della varianza-covarianza delle stime dei parametri nel modello in un nuovo file di dati della sessione attiva o in un file di dati SPSS Statistics esterno. Per ciascuna variabile dipendente è inoltre disponibile una riga di stime dei parametri, una riga di valori di significatività per le statistiche *t* corrispondenti alle stime dei parametri e una riga di gradi di libertà dei residui. Per ciascuna variabile dipendente di modelli multivariati sono disponibili righe simili. Il file matrice può essere usato in altre procedure che leggono i file matrice.

GLM – Univariato: Opzioni

Figura 11-8
Finestra di dialogo Opzioni



In questa finestra di dialogo sono disponibili statistiche opzionali. Le statistiche vengono calcolate tramite un modello di effetti fissi.

Medie marginali stimate. Selezionare i fattori e le interazioni per cui si desiderano le stime delle medie marginali della popolazione nelle celle. Queste medie vengono corrette per le eventuali covariate.

- **Confronta effetti principali.** Consente di eseguire confronti a coppie senza correzione tra le medie marginali stimate di qualsiasi effetto principale del modello, per fattori sia tra soggetti che entro soggetti. Questa opzione è disponibile solo se nell'elenco Medie marginali per sono stati selezionati effetti principali.
- **Correzione intervallo di confidenza.** Selezionare la differenza meno significativa (LSD), la correzione di Bonferroni o di Sidak agli intervalli di confidenza e la significatività. Questo comando è disponibile solo se è stato selezionato Confronta effetti principali.

Visualizzazione. Selezionare Statistiche descrittive per produrre medie osservate, deviazioni standard e conteggi per tutte le variabili dipendenti di tutte le celle. La funzione Stima della dimensione degli effetti fornisce un valore eta-quadrato parziale per ciascun effetto e per ciascuna stima dei parametri. La statistica eta-quadrato consente di ottenere la proporzione della variabilità totale attribuibile a un fattore. Selezionare Potenza osservata per ottenere la potenza del test nel caso in cui l'ipotesi alternativa sia basata sul valore osservato. Selezionare Stime dei parametri per ottenere

stime dei parametri, errori standard, test *t*, intervalli di confidenza e la potenza osservata per ciascun test. Selezionare Matrice dei coefficienti di contrasto per ottenere la matrice **L**.

La funzione Test di omogeneità produce il test di Levene per l'omogeneità della varianza per ogni variabile dipendente su tutte le combinazioni di livello dei fattori fra soggetti, solo per i fattori fra soggetti. Le opzioni Grafici di variabilità vs. densità e Grafici dei residui risultano utili per la verifica di ipotesi sui dati. Se non è disponibile alcun fattore, questa opzione risulta disattivata. Selezionare Grafici dei residui per ottenere un grafico dei residui osservati, attesi e standardizzati per ciascuna variabile dipendente. Questi grafici risultano utili per l'analisi dell'ipotesi di uguaglianza della varianza. Selezionare Mancanza di adattamento per controllare che la relazione fra la variabile dipendente e le variabili indipendenti possa essere descritta in modo adeguato dal modello. La funzione Forma funzionale generalizzata consente di creare test di ipotesi personalizzati basati sulla forma funzionale generalizzata. Le righe di una matrice dei coefficienti di contrasto sono combinazioni lineari della forma funzionale generalizzata.

Livello di significatività. Potrebbe risultare utile correggere il livello di significatività usato nei test post hoc e il livello di confidenza usato per la costruzione degli intervalli di confidenza. Il valore specificato viene inoltre usato per il calcolo della potenza osservata per il test. Quando si specifica un livello di significatività, nella finestra di dialogo viene visualizzato il livello di intervalli di confidenza associato.

Funzioni aggiuntive del comando UNIANOVA

Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare gli effetti nidificati del disegno (tramite il sottocomando `DESIGN`).
- Specificare test di effetti vs. una combinazione lineare di effetti o un valore (tramite il sottocomando `TEST`).
- Specificare contrasti multipli (tramite il sottocomando `CONTRAST`).
- Includere valori mancanti definiti dall'utente (tramite il sottocomando `MISSING`).
- Specificare criteri EPS (tramite il sottocomando `CRITERIA`).
- Costruire una matrice **L**, **M** o **K** personalizzata (tramite il sottocomando `LMATRIX`, `MMATRIX` o `KMATRIX`).
- Per i contrasti deviazione e i contrasti semplici, specificare una categoria di riferimento intermedia (tramite il sottocomando `CONTRAST`).
- Specificare metrica per contrasti polinomiali (tramite il sottocomando `CONTRAST`).
- Specificare termini di errore per confronti post-hoc (tramite il sottocomando `POSTHOC`).
- Calcolare medie marginali stimate per qualsiasi fattore o interazione tra fattori per i fattori elencati (tramite il sottocomando `EMMEANS`).
- Assegnare un nome alle variabili temporanee (tramite il sottocomando `SAVE`).
- Costruire un file di dati di matrici di correlazione (tramite il sottocomando `OUTFILE`).
- Costruire un file di dati di matrici contenente statistiche derivate dai dati della tabella ANOVA tra soggetti (tramite il sottocomando `OUTFILE`).
- Salvare la matrice del disegno in un nuovo file di dati (tramite il sottocomando `OUTFILE`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
12

Correlazioni bivariate

La procedura Correlazioni bivariate consente di calcolare coefficiente di correlazione di Pearson, rho di Spearman e tau-*b* di Kendall con i rispettivi livelli di significatività. Le correlazioni consentono di misurare la relazione tra variabili o punteggi di rango. Prima di calcolare un coefficiente di correlazione, è necessario valutare la presenza di valori anomali nei dati (che possono causare risultati errati) e l'esistenza di una relazione lineare. Il coefficiente di correlazione di Pearson è una misura di associazione lineare. Due variabili possono essere perfettamente correlate, ma se la relazione non è lineare il coefficiente di correlazione di Pearson non è la statistica migliore per misurare tale associazione.

Esempio. Il numero di partite vinte da una squadra di baseball è correlato con la media dei punti totalizzati per ciascuna partita? Un grafico a dispersione indica l'esistenza di una relazione lineare. Dall'analisi dei dati relativi alla stagione NBA 1994–1995 risulta che il coefficiente di correlazione di Pearson (0,581) è significativo al livello 0,01. Si può presumere che il numero di partite vinte per stagione sia inversamente proporzionale ai punti totalizzati dagli avversari. Queste variabili sono legate da una correlazione negativa (–0,401), significativa al livello 0,05.

Statistiche. Per ogni variabile: numero di casi con valori non mancanti, media e deviazione standard. Per ciascuna coppia di variabili: coefficiente di correlazione di Pearson, rho di Spearman, tau-*b* di Kendall, prodotto incrociato delle deviazioni, covarianza.

Dati. Utilizzare le variabili quantitative simmetriche per il coefficiente di correlazione di Pearson e le variabili quantitative o le variabili con categorie ordinate per rho di Spearman e tau-*b* di Kendall.

Assunzioni. Il coefficiente di correlazione di Pearson assume che ciascuna coppia di variabili sia bivariata normale.

Per ottenere correlazioni bivariate

Dai menu, scegliere:
Analizza > Correlazione > Bivariata...

Figura 12-1
Finestra di dialogo Correlazioni bivariate

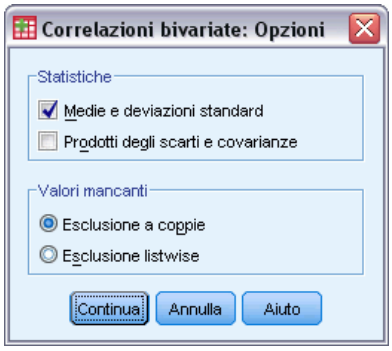
▶ Selezionare una o più variabili numeriche.

Sono inoltre disponibili le seguenti opzioni:

- **Coefficienti di correlazione.** Per variabili quantitative normalmente distribuite, scegliere il coefficiente di correlazione di Pearson. Se i dati non sono normalmente distribuiti o prevedono categorie ordinate, scegliere Tau-b di Kendall o Spearman, per misurare l'associazione tra punteggi di rango. I valori dei coefficienti di correlazione variano da –1 (relazione negativa perfetta) a +1 (relazione positiva perfetta). Il valore 0 indica l'assenza di relazione lineare. Interpretando i risultati, evitare di trarre conclusioni di tipo causa-effetto sulla base di una correlazione significativa.
- **Test di significatività.** È possibile selezionare le probabilità a una coda o a due code. Se si conosce in anticipo la direzione dell'associazione, selezionare A una coda.. In alternativa, selezionare A due code.
- **Evidenzia correlazioni significative.** I coefficienti di correlazione significativi al livello 0,05 vengono identificati con un asterisco singolo e quelli significativi al livello 0,01 con due asterischi.

Correlazioni bivariate: Opzioni

Figura 12-2
Finestra di dialogo Correlazioni bivariate: Opzioni



Statistiche. Per le correlazioni di Pearson, è possibile scegliere una o entrambe le seguenti opzioni.

- **Medie e deviazioni standard.** Visualizzate per ciascuna variabile. Viene indicato anche il numero di casi con valori non mancanti. I valori mancanti vengono gestiti variabile per variabile indipendentemente dall'impostazione corrispondente.
- **Prodotti degli scarti e delle covarianze.** Viene visualizzato per ciascuna coppia di variabili. La deviazione del prodotto incrociato è equivalente alla somma dei prodotti delle variabili corrette per la media. È il numeratore del coefficiente di correlazione di Pearson. La covarianza è una misura non standardizzata della relazione tra due variabili, equivalente alla deviazione del prodotto incrociato divisa per *N*–1.

Valori mancanti. È possibile scegliere tra le opzioni seguenti:

- **Esclusione pairwise.** I casi con valori mancanti per una o entrambe le variabili di una coppia per un coefficiente di correlazione vengono esclusi dall'analisi. Poiché ciascun coefficiente si basa su tutti i casi con codici validi per quella particolare coppia di variabili, in tutti i calcoli verrà utilizzato il maggior numero di informazioni disponibile. In questo modo è possibile ottenere una serie di coefficienti basati su un numero variabile di casi.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile vengono esclusi da tutte le correlazioni.

Funzioni aggiuntive dei comandi CORRELATIONS e NONPAR CORR

Il linguaggio della sintassi dei comandi consente inoltre di:

- Scrivere una matrice di correlazione per la correlazione di Pearson da utilizzare in luogo dei dati per ottenere altri tipi di analisi, ad esempio l'analisi fattoriale (con il sottocomando `MATRIX`).
- Ottenere correlazioni di ciascuna variabile presente in un elenco con le variabili corrispondenti presenti in un secondo elenco (utilizzando la parola chiave `WITH` con il sottocomando `VARIABLES`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
13

Correlazioni parziali

La procedura Correlazioni parziali consente di calcolare i coefficienti di correlazione parziale che descrivono la relazione lineare tra due variabili controllando gli effetti di una o più variabili aggiuntive. Le correlazioni sono misure di associazione lineare. Due variabili possono essere perfettamente correlate, ma se la relazione non è lineare, un coefficiente di correlazione non rappresenta la statistica più adatta a misurarne l'associazione.

Esempio. Esiste una correlazione tra i fondi stanziati per il sistema sanitario e il tasso di malattie? Sebbene si potrebbe pensare che una relazione di tale tipo sia negativa, da uno studio emerge una correlazione *positiva* significativa: con l'aumentare dei fondi stanziati per il sistema sanitario, il tasso di malattie sembra aumentare. Il controllo della frequenza delle visite ai medici elimina virtualmente la correlazione positiva osservata. I fondi stanziati per il sistema sanitario e il tasso di malattie sembrano avere una correlazione positiva solo perché più pazienti hanno accesso al sistema sanitario quando vengono aumentati i fondi, con un conseguente aumento delle malattie segnalate da medici e ospedali.

Statistiche. Per ogni variabile: numero di casi con valori non mancanti, media e deviazione standard. Matrici di correlazione parziale e di ordine zero, con gradi di libertà e livelli di significatività.

Dati. Utilizzare variabili simmetriche, quantitative.

Assunzioni. La procedura Correlazioni parziali si fonda sull'ipotesi che ciascuna coppia di variabili sia bivariata normale.

Per ottenere correlazioni parziali

- Dai menu, scegliere:
  Analizza > Correlazione > Parziale...

Figura 13-1
Finestra di dialogo Correlazioni parziali

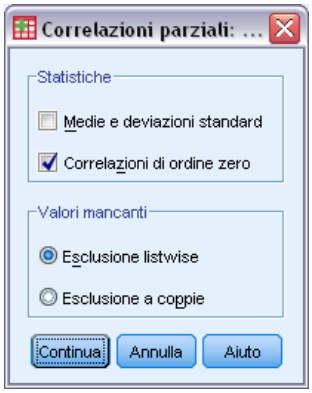
- Selezionare due o più variabili numeriche per cui è necessario calcolare le correlazioni parziali.
- Selezionare una o più variabili numeriche di controllo.

Sono inoltre disponibili le seguenti opzioni:

- **Test di significatività.** È possibile selezionare le probabilità a una coda o a due code. Se si conosce in anticipo la direzione dell'associazione, selezionare A una coda.. In alternativa, selezionare A due code.
- **Mostra l'esatto livello di significatività.** Per impostazione predefinita, vengono indicati la probabilità e i gradi di libertà di ciascun coefficiente di correlazione. Se questa opzione viene deselezionata, i coefficienti significativi al livello 0,05 vengono identificati con un solo asterisco, i coefficienti significativi al livello 0,01 con un doppio asterisco e i gradi di libertà vengono eliminati. Questa impostazione viene applicata alle matrici di correlazione parziale e di ordine zero.

Correlazioni parziali: Opzioni

Figura 13-2
Finestra di dialogo Correlazioni parziali: Opzioni



Statistiche. È possibile scegliere una delle seguenti opzioni o entrambe:

- **Medie e deviazioni standard.** Visualizzate per ciascuna variabile. Viene indicato anche il numero di casi con valori non mancanti.
- **Correlazioni di ordine zero.** Viene visualizzata una matrice di correlazioni semplici tra tutte le variabili, incluse le variabili di controllo.

Valori mancanti. È possibile scegliere una delle seguenti opzioni:

- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile, incluse le variabili di controllo, vengono esclusi da tutti i calcoli.
- **Esclusione pairwise.** Per il calcolo delle correlazioni di ordine zero su cui si basano le correlazioni parziali, i casi con valori mancanti per una o entrambe le variabili di una coppia non verranno utilizzati. L'eliminazione pairwise consente di utilizzare il massimo numero di dati possibile. Il numero di casi, tuttavia, può differire a seconda del coefficiente. Quando è attiva l'eliminazione pairwise, i gradi di libertà per un particolare coefficiente parziale si basano sul numero minimo di casi utilizzati per il calcolo di una delle correlazioni di ordine zero.

Opzioni aggiuntive del comando PARTIAL CORR

Il linguaggio della sintassi dei comandi consente inoltre di:

- Leggere una matrice di correlazione di ordine zero o scrivere una matrice di correlazione parziale (con il sottocomando `MATRIX`).
- Ottenere correlazioni parziali tra due elenchi di variabili (con la parola chiave `WITH` nel sottocomando `VARIABLES`).
- Ottenere più analisi (con più sottocomandi `VARIABLES`).
- Specificare i valori degli ordini da richiedere (ad esempio sia le correlazioni parziali di primo e secondo ordine) quando sono disponibili due variabili di controllo (con il sottocomando `VARIABLES`).
- Eliminare i coefficienti ridondanti (con il sottocomando `FORMAT`).
- Visualizzare una matrice di correlazioni semplici quando non è possibile calcolare alcuni coefficienti (con il sottocomando `STATISTICS`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
14

Distanze

Questa procedura consente di calcolare una grande varietà di statistiche in base alle similarità o alle dissimilarità (distanze), sia considerando coppie di variabili, sia considerando coppie di casi. Tali misure di distanza o similarità possono poi essere utilizzate insieme ad altre procedure quali analisi fattoriale, analisi cluster o scaling multidimensionale per l'analisi di insiemi complessi di dati.

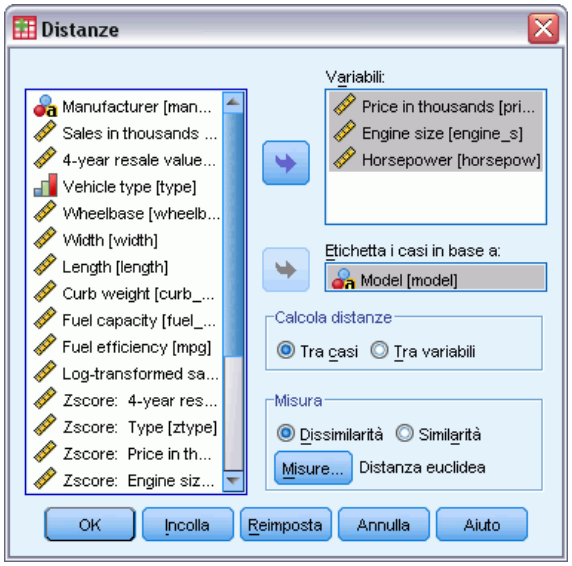
Esempio. È possibile misurare le similarità tra coppie di automobili in base a caratteristiche specifiche, quali dimensione del motore, MPG e potenza. Grazie al calcolo delle similarità, è possibile stabilire quali automobili sono simili e quali sono differenti tra loro. Per un'analisi più formale, si può scegliere di applicare alle similarità l'analisi cluster gerarchica o lo scaling multidimensionale che consentono di esplorare la struttura sottostante.

Statistiche. Le misure di dissimilarità (distanza) disponibili per i dati di intervallo sono: distanza euclidea, distanza euclidea quadratica, Chebychev, City-block, Minkowski, personalizzata. Per i dati di conteggio, le misure disponibili sono chi-quadrato o phi-quadrato. Per i dati binari, infine, le misure disponibili sono: distanza euclidea, distanza euclidea quadratica, differenza di dimensione, differenza di modello, varianza, forma, Lance e Williams. Le misure di similarità disponibili per i dati di intervallo sono la correlazione di Pearson o il coseno. Per i dati binari sono invece disponibili le seguenti misure: Russel e Rao, corrispondenza semplice, Jaccard, Dice, Rogers e Tanimoto, Sokal e Sneath 1, Sokal e Sneath 2, Sokal e Sneath 3, Kulczynski 1, Kulczynski 2, Sokal e Sneath 4, Hamann, Lambda, *D* di Anderberg, *Y* di Yule, *Q* di Yule, Ochiai, Sokal e Sneath 5, correlazione phi a 4 punti o dispersione.

Per ottenere matrici delle distanze

- Dai menu, scegliere:
  Analizza > Correlazione > Distanze...

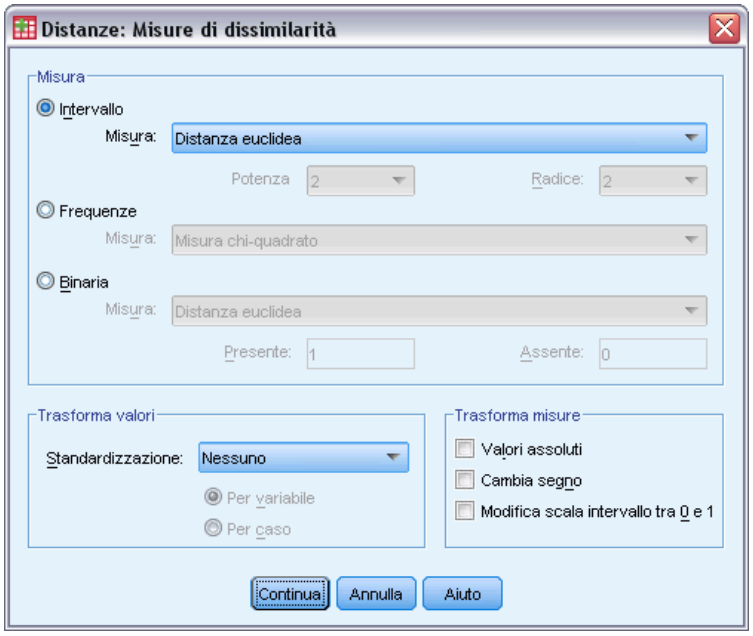
Figura 14-1
Finestra di dialogo Distanze



- Selezionare una variabile numerica per calcolare la distanza tra casi e almeno due variabili numeriche per calcolare la distanza tra variabili.
- Scegliere un'alternativa nel gruppo Calcola distanze per calcolare le distanze tra casi o tra variabili.

Distanze: Misure di dissimilarità

Figura 14-2
Finestra di dialogo Distanze: Misure di dissimilarità



Dal gruppo Misura selezionare l'alternativa corrispondente al tipo di dati desiderato (intervallo, conteggio o binari), quindi scegliere dall'elenco a discesa una delle misure corrispondenti a quel tipo di dati. Le misure disponibili per tipo di dati sono le seguenti:

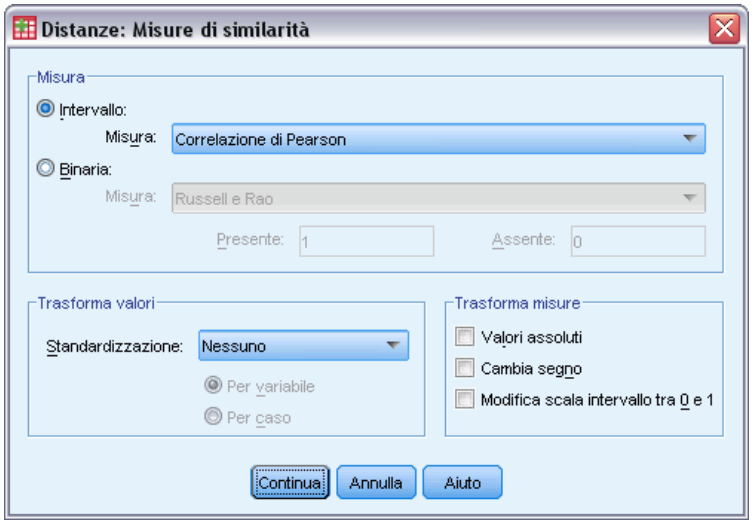
- **Dati per intervallo.** Distanza euclidea, distanza euclidea quadratica, Chebychev, City-block, Minkowski o personalizzata.
- **Conteggio dati.** Misura chi-quadrato e misura phi-quadrato.
- **Dati binari.** Distanza euclidea, distanza euclidea quadratica, differenza di dimensione, differenza di modello, varianza, forma o di Lance e Williams. (Inserire i valori Presente e Assente per specificare i due valori significativi, tutti gli altri valori verranno ignorati dalle distanze).

Il gruppo Trasforma valori consente di standardizzare i valori per casi o valori *prima* di calcolare le similarità. Tali trasformazioni non sono applicabili ai dati binari. I metodi di standardizzazione disponibili sono: punteggi *z*, intervallo da –1 a 1, intervallo da 0 a 1, ampiezza massima di 1, media di 1 o deviazione standard di 1.

Il gruppo Trasforma misure consente di trasformare i valori generati dalla misura della distanza. Questi verranno applicati dopo il calcolo della misura di distanza. Le alternative disponibili sono: Valori assoluti, Cambia segno e Riscala all'intervallo tra 0 e 1.

Distanze: Misure di similarità

Figura 14-3
Finestra di dialogo Distanze: Misure di similarità



Selezionare l'alternativa corrispondente al tipo di dati desiderato (intervallo o binari) dal gruppo Misura , quindi scegliere una delle misure corrispondenti a quel tipo di dati dall'elenco a discesa. Le misure disponibili per tipo di dati sono le seguenti:

- **Dati per intervallo.** Correlazione Pearson o coseno.
- **Dati binari.** Russel e Rao, corrispondenza semplice, Jaccard, Dice, Rogers e Tanimoto, Sokal e Sneath 1, Sokal e Sneath 2, Sokal e Sneath 3, Kulczynski 1, Kulczynski 2, Sokal e Sneath 4, Hamann, Lambda, *D* di Anderberg, *Y* di Yule, *Q* di Yule, Ochiai, Sokal e Sneath 5,

correlazione phi a 4 punti o dispersione. (Inserire i valori Presente e Assente per specificare i due valori significativi, tutti gli altri valori verranno ignorati dalle distanze).

Il gruppo Trasforma valori consente di standardizzare i valori per casi o valori prima di calcolare le similarità. Tali trasformazioni non sono applicabili ai dati binari. I metodi di standardizzazione disponibili sono: punteggi *z*, intervallo da –1 a 1, intervallo da 0 a 1, ampiezza massima di 1, media di 1 e deviazione standard di 1.

Il gruppo Trasforma misure consente di trasformare i valori generati dalla misura della distanza. Questi verranno applicati dopo il calcolo della misura di distanza. Le alternative disponibili sono: Valori assoluti, Cambia segno e Riscala all'intervallo tra 0 e 1.

Opzioni aggiuntive del comando PROXIMITIES

La procedura Distanze usa la sintassi del comando `PROXIMITIES`. Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare un numero intero come potenza per la misura della distanza di Minkowski.
- Specificare qualsiasi intero come potenza e radice per la misura personalizzata delle distanze.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
15

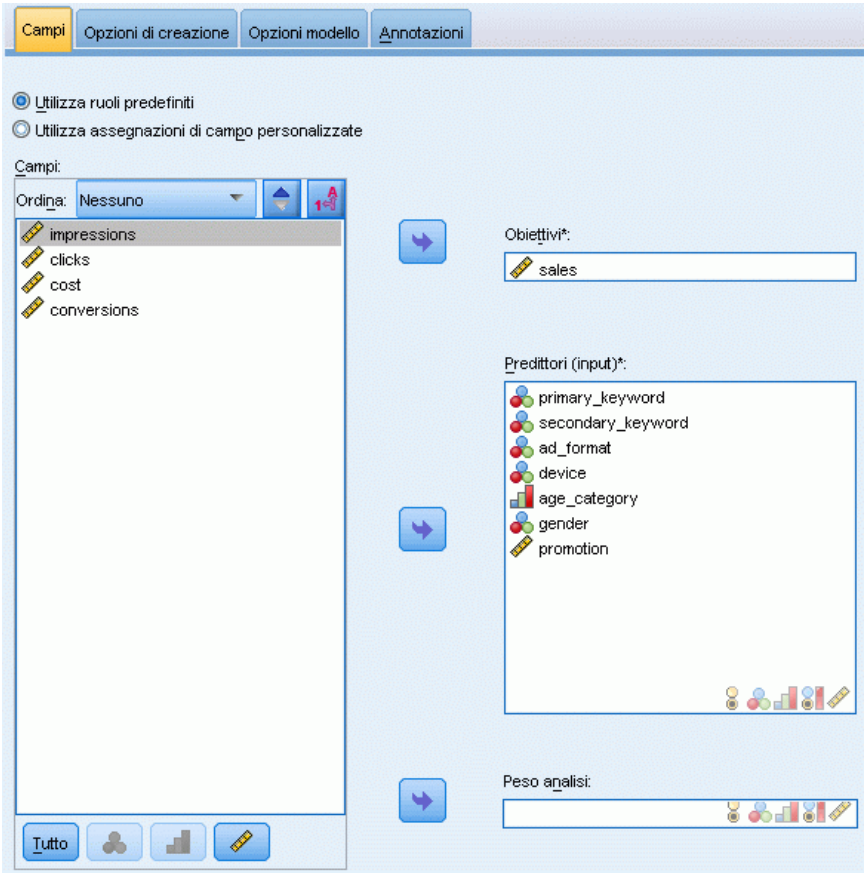
Modelli lineari

I modelli lineari prevedono un obiettivo continuo basato sulle relazioni lineari tra l'obiettivo e uno o più predittori.

I modelli lineari sono relativamente semplici e forniscono una formula matematica di facile interpretazione per il calcolo del punteggio. Le proprietà di questi modelli sono di facile comprensione e di solito possono essere generate molto rapidamente rispetto ad altri tipi di modelli, quali le reti neurali o gli alberi decisionali, nello stesso insieme di dati.

Esempio. Una compagnia di assicurazioni con poche risorse per indagare sulle richieste di indennizzo dei proprietari immobiliari vuole creare un modello per stimare i costi di tali richieste di indennizzo. Tramite l'implementazione di questo modello nei centri di servizio, i rappresentanti possono inserire le informazioni relative alla richiesta di indennizzo mentre sono al telefono con un cliente e ottenere immediatamente il costo "previsto" della richiesta in base ai dati passati.

Figura 15-1
Scheda Campi



Requisiti dei campi. Devono esservi un campo Obiettivo e almeno un campo Input. Per impostazione predefinita, i campi con ruoli predefiniti Entrambi o Nessuno non vengono utilizzati. L'obiettivo deve essere continuo (scala). Non vi sono restrizioni per il livello di misurazione sui predittori (input); i campi categoriali (nominali e ordinali) sono utilizzati come fattori nel modello e i campi continui sono utilizzati come covariate.

Nota: se un campo categoriale contiene più di 100 categorie, la procedura non viene eseguita e non viene creato alcun modello.

Per ottenere un modello lineare

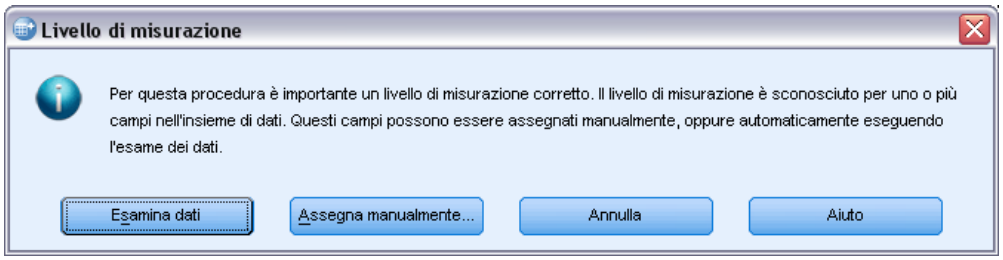
Questa funzione richiede il modulo Statistics Base.

Dai menu, scegliere:
Analizza > Regression > Modellazione lineare automatica

- Assicurarsi che siano presenti almeno un obiettivo e un input.
- Fare clic su Opzioni di creazione per specificare le impostazioni facoltative per la creazione e il modello.
- Fare clic su Opzioni modello per salvare i punteggi negli insiemi di dati attivi ed esportare il modello in un file esterno.
- Fare clic su Esegui per eseguire la procedura e creare gli oggetti Modello.

L'avviso Livello di misurazione viene visualizzato quando il livello di misurazione di una o più variabili (campi) dell'insieme di dati è sconosciuto. Poiché influisce sul calcolo dei risultati di questa procedura, il livello di misurazione deve essere definito per tutte le variabili.

Figura 15-2
Avviso Livello di misurazione



- **Esamina dati.** Legge i dati dell'insieme di dati attivo e assegna un livello di misurazione predefinito a tutti i campi con livello di misurazione sconosciuto. Con insiemi di dati di grandi dimensioni, questa operazione può richiedere del tempo.
- **Assegna manualmente.** Apre una finestra di dialogo che elenca tutti i campi con livello di misurazione sconosciuto, mediante la quale è possibile assegnare un livello di misurazione a questi campi. Il livello di misurazione si può assegnare anche nella Visualizzazione variabili dell'Editor dei dati.

Dal momento che il livello di misurazione è importante per questa procedura, è possibile accedere alla finestra di dialogo per la sua esecuzione solo quando per tutti i campi è stato definito un livello di misurazione.

Obiettivi

Qual è l'obiettivo principale?

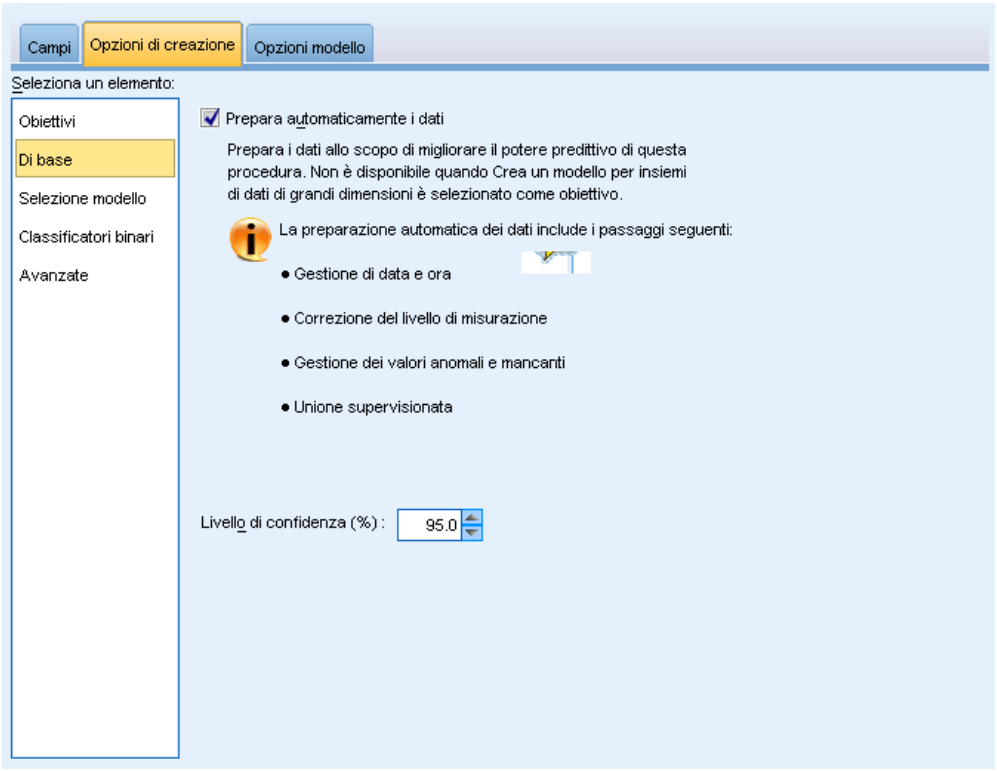
- **Crea un modello standard.** Il metodo crea un singolo modello per prevedere l'obiettivo utilizzando i predittori. In genere i modelli standard sono più semplici da interpretare e il calcolo del punteggio può risultare più rapido rispetto ai classificatori binari di cui è stato eseguito il boosting, il bagging o ai classificatori binari di insiemi di dati di grandi dimensioni.
- **Migliora la precisione del modello (boosting).** Il metodo crea un modello Classificatore binario tramite boosting, che genera una sequenza di modelli per ottenere previsioni più precise. La creazione e il calcolo del punteggio dei classificatori binari può richiedere più tempo rispetto ai modelli standard.

  Il boosting produce una successione di "modelli di componente", ognuno dei quali viene costruito sull'insieme di dati intero. Prima di costruire ogni modello di componente successivo, i record vengono ponderati sulla base dei residui del modello di componente precedente. Ai casi con residui considerevoli vengono assegnati pesi di analisi relativamente più elevati in modo che il modello di componente successivo si concentri sulla previsione accurata di quei record. Insieme questi modelli di componente formano un modello Classificatore binario. Il modello Classificatore binario calcola il punteggio dei nuovi record utilizzando una regola di combinazione; le regole disponibili dipendono dal livello di misurazione dell'obiettivo.
- **Migliora la stabilità del modello (bagging).** Il metodo crea un modello Classificatore binario tramite bagging (aggregazione bootstrap), che genera più modelli per ottenere previsioni più affidabili. La creazione e il calcolo del punteggio dei classificatori binari può richiedere più tempo rispetto ai modelli standard.

  L'aggregazione bootstrap (bagging) produce delle repliche dell'insieme di dati di training mediante campionamento con sostituzione dall'insieme di dati originale. In questo modo vengono creati campioni di bootstrap di dimensioni uguali all'insieme di dati originale. Quindi viene costruito un "modello di componente" per ogni replica. Insieme questi modelli di componente formano un modello Classificatore binario. Il modello Classificatore binario calcola il punteggio dei nuovi record utilizzando una regola di combinazione; le regole disponibili dipendono dal livello di misurazione dell'obiettivo.
- **Crea un modello per insiemi di dati di grandi dimensioni (richiede IBM® SPSS® Statistics Server).** Il metodo crea un modello Classificatore binario suddividendo l'insieme di dati in blocchi di dati. Scegliere questa opzione se l'insieme di dati è troppo grande per creare uno dei modelli descritti oppure per la creazione di modelli incrementali. Questa opzione risulta più rapida per la creazione, ma può richiedere più tempo per il calcolo del punteggio rispetto a un modello standard. Questa opzione richiede la connettività SPSS Statistics Server.

Opzioni di base

Figura 15-3
Impostazioni di base



Prepara automaticamente i dati. Questa opzione consente di trasformare internamente l'obiettivo e i predittori in modo tale da ottimizzare il potere predittivo del modello; eventuali trasformazioni vengono salvate con il modello e applicate ai nuovi dati per il calcolo del punteggio. Le versioni originali dei campi trasformati vengono escluse dal modello. Per default, vengono eseguite le seguenti operazioni di preparazione automatica dei dati.

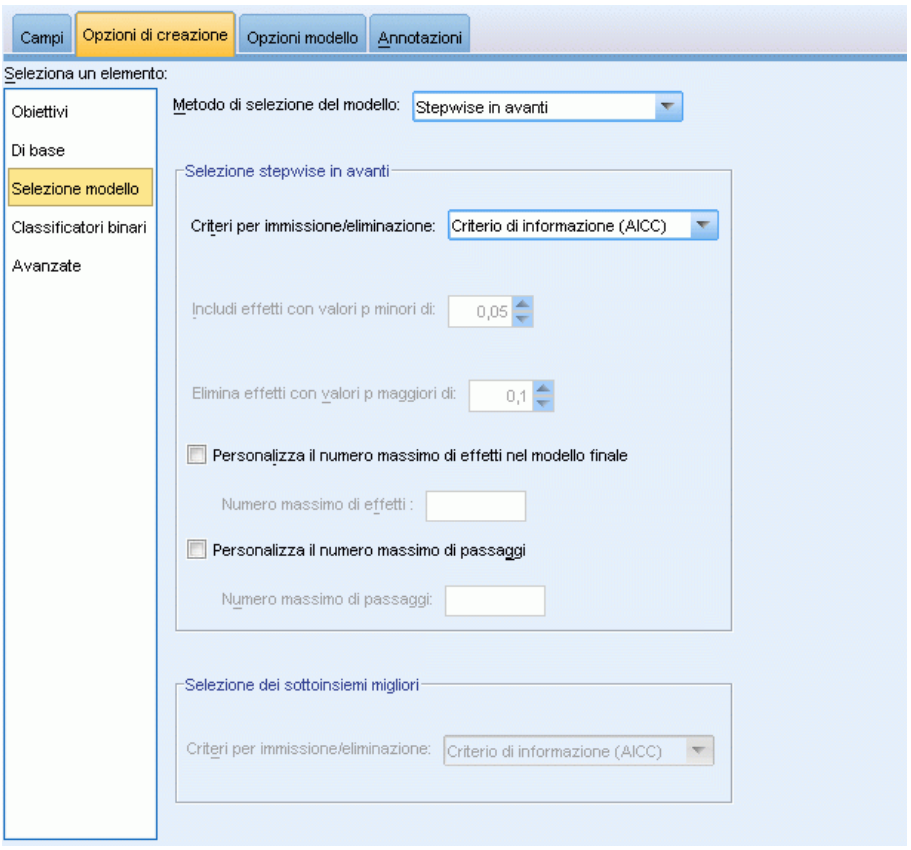
- **Gestione di data e ora.** Ogni predittore di data viene trasformato in un nuovo predittore continuo contenente il tempo trascorso a partire da una data di riferimento (01/01/1970). Ogni predittore di ora viene trasformato in un nuovo predittore continuo contenente il tempo trascorso a partire da un orario di riferimento (00:00:00).
- **Regola livello di misurazione.** I predittori continui con meno di 5 valori distinti vengono riformulati come predittori ordinali. I predittori ordinali con più di 10 valori distinti vengono riformulati come predittori continui.
- **Gestione dei valori anomali.** I valori dei predittori continui che si posizionano oltre un valore di interruzione (3 deviazioni standard dalla media) vengono impostati sul valore di interruzione.

- **Gestione dei valori mancanti.** I valori mancanti dei predittori nominali vengono sostituiti con la modalità della partizione di addestramento. I valori mancanti dei predittori ordinali vengono sostituiti con la mediana della partizione di addestramento. I valori mancanti dei predittori continui vengono sostituiti con la media della partizione di addestramento.
- **Unione supervisionata.** Questa opzione consente di creare un modello più gestibile riducendo il numero dei campi da elaborare in associazione all'obiettivo. Le categorie simili vengono identificate in base alla relazione tra input e obiettivo. Le categorie che non presentano differenze significative (ovvero che hanno un valore p superiore a 0,1) vengono unite. Se tutte le categorie vengono unite in una, la versione originale e quella derivata del campo vengono escluse dal modello perché non hanno un valore come predittore.

Livello di confidenza. Si tratta del livello di confidenza utilizzato per calcolare stime di intervallo per i coefficienti del modello nella visualizzazione Coefficienti. Specificare un valore maggiore di 0 e minore di 100. L'impostazione di default è 95.

Selezione del modello

Figura 15-4
Impostazioni della selezione del modello



Metodo di selezione del modello. Scegliere uno dei metodi di selezione del modello descritti di seguito o Includi tutti i predittori, che immette tutti i predittori disponibili come termini del modello degli effetti principali. Per default si utilizza il metodo Stepwise in avanti.

Selezione Stepwise in avanti. La selezione viene avviata senza effetti nel modello e aggiunge o elimina effetti un passaggio alla volta finché non vi sono più effetti da aggiungere o eliminare, in base ai criteri stepwise.

- **Criteri per immissione/eliminazione.** Statistica utilizzata per determinare se un effetto deve essere aggiunto o eliminato dal modello. Il criterio di informazione (AICC) si basa sulla probabilità dell'insieme di addestramento in relazione al modello e viene adeguato in modo da penalizzare i modelli eccessivamente complessi. Statistiche F si basa su un test statistico del miglioramento nell'errore del modello. R-quadrato corretto si basa sull'adattamento dell'insieme di addestramento e viene adeguato in modo da penalizzare i modelli eccessivamente complessi. Il criterio di prevenzione del sovradattamento (ASE) si basa sull'adattamento (errore quadratico medio, o ASE) dell'insieme di prevenzione del sovradattamento. L'insieme di prevenzione del sovradattamento è un sottocampione di circa il 30% dell'insieme di dati originale che non viene utilizzato per addestrare il modello.

  Se si sceglie un criterio diverso da Statistiche F, in ogni passaggio l'effetto che corrisponde al massimo aumento positivo nel criterio viene aggiunto al modello. Tutti gli effetti presenti nel modello che corrispondono a una diminuzione nel criterio vengono eliminati.

  Se si sceglie il criterio Statistiche F, in ogni passaggio l'effetto che ha il valore *p* più piccolo inferiore alla soglia specificata, Includi effetti con valori p minori di, viene aggiunto al modello. Il valore di default è 0,05. Tutti gli effetti presenti nel modello che hanno un valore *p* superiore alla soglia specificata, Elimina effetti con valori p maggiori di, vengono eliminati. Il valore di default è 0,10.

- **Personalizza il numero massimo di effetti nel modello finale.** Per default, tutti gli effetti disponibili possono essere immessi nel modello. In alternativa, se l'algoritmo stepwise conclude un passaggio con il numero massimo di effetti specificato, l'algoritmo si interrompe in corrispondenza dell'insieme di effetti corrente.
- **Personalizza il numero massimo di passaggi.** L'algoritmo stepwise si interrompe dopo un determinato numero di passaggi. Per default, il numero è 3 volte il numero di effetti disponibili. In alternativa, specificare un numero intero positivo come numero massimo di passaggi.

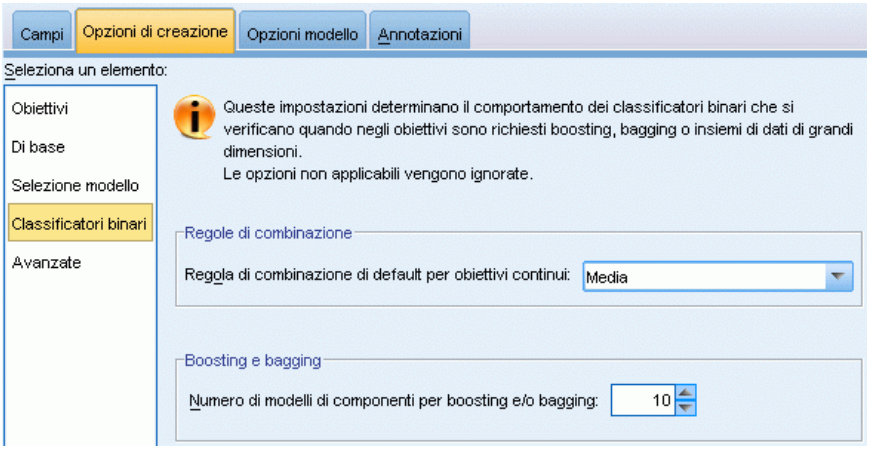
Selezione dei sottoinsiemi migliori. Verifica tutti i modelli "possibili" o almeno un sottoinsieme di modelli possibili più grande rispetto al metodo stepwise in avanti, in modo tale da scegliere il migliore in assoluto in base al criterio di selezione dei sottoinsiemi migliori. Il criterio di informazione (AICC) si basa sulla probabilità dell'insieme di addestramento in relazione al modello e viene adeguato in modo da penalizzare i modelli eccessivamente complessi. R-quadrato corretto si basa sull'adattamento dell'insieme di addestramento e viene adeguato in modo da penalizzare i modelli eccessivamente complessi. Il criterio di prevenzione del sovradattamento (ASE) si basa sull'adattamento (errore quadratico medio, o ASE) dell'insieme di prevenzione del sovradattamento. L'insieme di prevenzione del sovradattamento è un sottocampione di circa il 30% dell'insieme di dati originale che non viene utilizzato per addestrare il modello.

Il modello con il massimo valore del criterio viene scelto come modello migliore.

Nota: La selezione dei sottoinsiemi migliori è maggiormente impegnativa a livello di calcolo rispetto alla selezione stepwise in avanti. Quando la sezione dei sottoinsiemi migliori viene eseguita in associazione a boosting, bagging o insiemi di dati di grandi dimensioni, la creazione può richiedere molto più tempo rispetto alla creazione di un modello standard con la selezione stepwise in avanti.

Classificatori binari

Figura 15-5
Impostazioni dei classificatori binari



Queste impostazioni determinano il comportamento dei classificatori binari che si verificano quando negli obiettivi sono richiesti boosting, bagging o insiemi di dati di grandi dimensioni. Le opzioni non applicate all'obiettivo selezionato vengono ignorate.

Bagging e insiemi di dati di grandi dimensioni. Quando si calcola il punteggio di un classificatore binario, questo tipo di regola consente di combinare i valori previsti provenienti dai modelli di base per calcolare il valore del punteggio del classificatore binario.

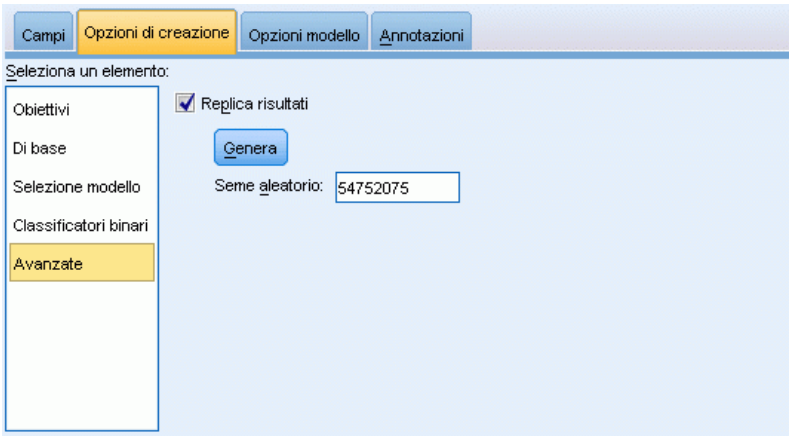
- **Regola di combinazione di default per obiettivi continui.** I valori previsti dei classificatori binari per gli obiettivi continui possono essere combinati utilizzando la media o la mediana dei valori previsti ricavati dai modelli di base.

Si noti che quando l'obiettivo è di ottimizzare la precisione del modello, le selezioni delle regole di combinazione vengono ignorate. Nel boosting viene sempre utilizzato un voto di maggioranza pesato per il calcolo del punteggio degli obiettivi categoriali e una mediana pesata per il calcolo del punteggio degli obiettivi continui.

Boosting e bagging. Specificare il numero dei modelli di base da creare quando l'obiettivo è di ottimizzare la precisione o la stabilità del modello. Per il bagging, si tratta del numero di campioni di bootstrap. Questo valore deve essere un numero intero positivo.

Opzioni avanzate

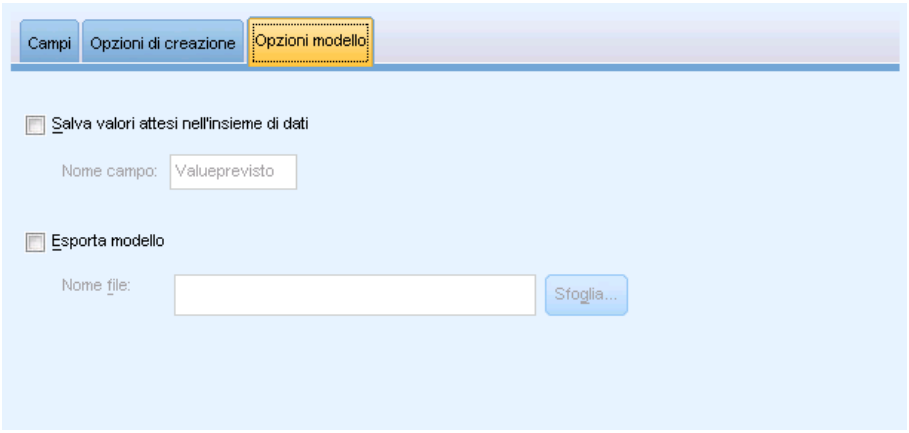
Figura 15-6
Impostazioni avanzate



Replica risultati. L'impostazione di un seme aleatorio consente di replicare le analisi. Il generatore di numeri casuali viene utilizzato per scegliere i record presenti nell'insieme di prevenzione del sovradattamento. Specificare un intero o fare clic su Genera per creare un intero pseudocasuale compreso tra 1 e 2147483647 incluso. Il valore di default è 54752075.

Opzioni modello

Figura 15-7
Scheda Opzioni modello

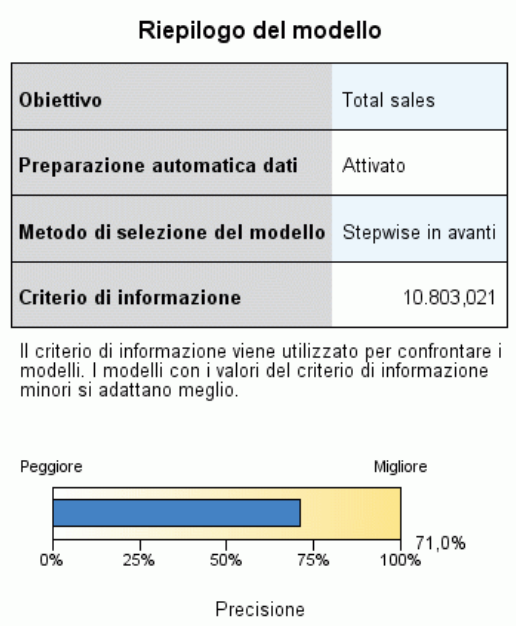


Salvare i valori previsti nell'insieme di dati. Il nome di default della variabile è *PredictedValue*.

Esporta modello. Consente di scrivere il modello in un file *.zip* esterno. È possibile utilizzare questo file di modello per applicare le informazioni del modello ad altri file di dati per il calcolo del punteggio. Specificare un nome file valido e univoco. Se la definizione del file fa riferimento a un file esistente, il file verrà sovrascritto.

Riepilogo del modello

Figura 15-8
Visualizzazione Riepilogo modello



La visualizzazione Riepilogo modello è un'istantanea, un riepilogo immediato del modello.

Tabella. La tabella identifica alcune impostazioni dei modelli di alto livello, inclusi:

- Il nome dell'obiettivo specificato nella scheda Campi,
- Se è stata eseguita la preparazione automatica dei dati come specificato nelle impostazioni Di base,
- Il metodo di selezione del modello e il criterio di selezione specificati nelle impostazioni Selezione modello. Viene anche visualizzato il valore del criterio di selezione per il modello finale, presentato in un formato ridotto.

Grafico. Il grafico visualizza la precisione del modello finale, presentata in un formato ingrandito. Il valore è 100 × il valore R^2 regolato per il modello finale.

Preparazione automatica dati

Figura 15-9
Visualizzazione Preparazione automatica dati

Preparazione automatica dati

Obiettivo: Total sales

Campo	Ruolo	Azioni intraprese
Age category	Predittore	Unisci categorie per aumentare al massimo l"associazione all"obiettivo
Primary keyword set	Predittore	Unisci categorie per aumentare al massimo l"associazione all"obiettivo
Promotion	Predittore	Cambia il livello di misurazione da continuo a ordinale
Secondary keyword set	Predittore	Unisci categorie per aumentare al massimo l"associazione all"obiettivo

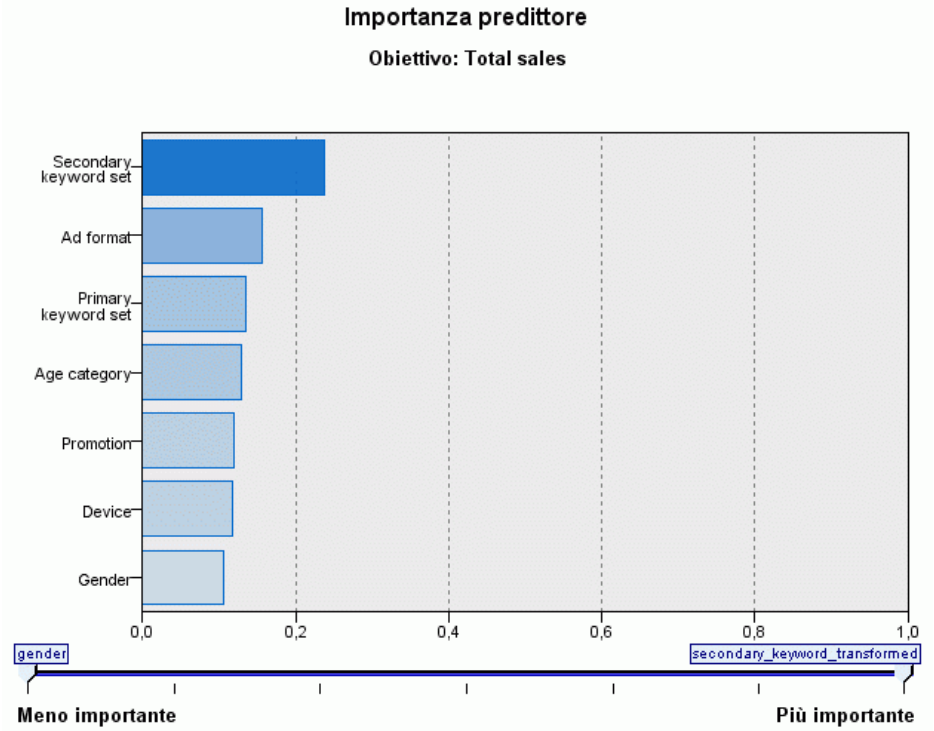
Se il nome del campo originale è X, il nome del campo trasformato è X_trasformato. Il campo originale viene escluso dall"analisi e al suo posto viene incluso il campo trasformato.

Questa visualizzazione contiene informazioni sui campi esclusi e sulla modalità di derivazione dei campi trasformati nel passaggio di preparazione automatica dei dati (ADP). Per ogni campo trasformato o escluso, la tabella indica il nome del campo, il ruolo nell'analisi e l'azione intrapresa nel passaggio dell'ADP. I campi sono ordinati in ordine alfabetico crescente in base al nome. Le possibili azioni adottate per ogni campo includono:

- Deriva durata: mesi calcola il tempo trascorso, in mesi, dai valori di un campo contenente le date fino alla data corrente del sistema.
- Deriva durata: ore calcola il tempo trascorso, in ore, dai valori di un campo contenente le ore fino all'ora corrente del sistema.
- Cambia il livello di misurazione da continuo a ordinale riformula i campi continui con meno di 5 valori univoci come campi ordinali.
- Cambia il livello di misurazione da ordinale a continuo riformula i campi ordinali con meno di 10 valori univoci come campi continui.
- Taglia valori anomali imposta i valori dei predittori continui che si posizionano oltre un valore di interruzione (3 deviazioni standard dalla media) sul valore di interruzione.
- Sostituisci valori mancanti sostituisce i valori mancanti dei campi nominali con la moda, i campi ordinali con la mediana e i campi continui con la media.
- Unisci categorie per aumentare al massimo l'associazione all'obiettivo identifica categorie di predittori "simili" in base alla relazione tra input e obiettivo. Le categorie che non presentano differenze significative (ovvero che hanno un valore *p* superiore a 0,05) vengono unite.
- Escludi predittore costante / dopo la gestione dei valori anomali / dopo l'unione delle categorie rimuove i predittori che hanno un solo valore, possibilmente dopo aver adottato altre azioni di preparazione automatica dei dati (ADP).

Importanza predittore

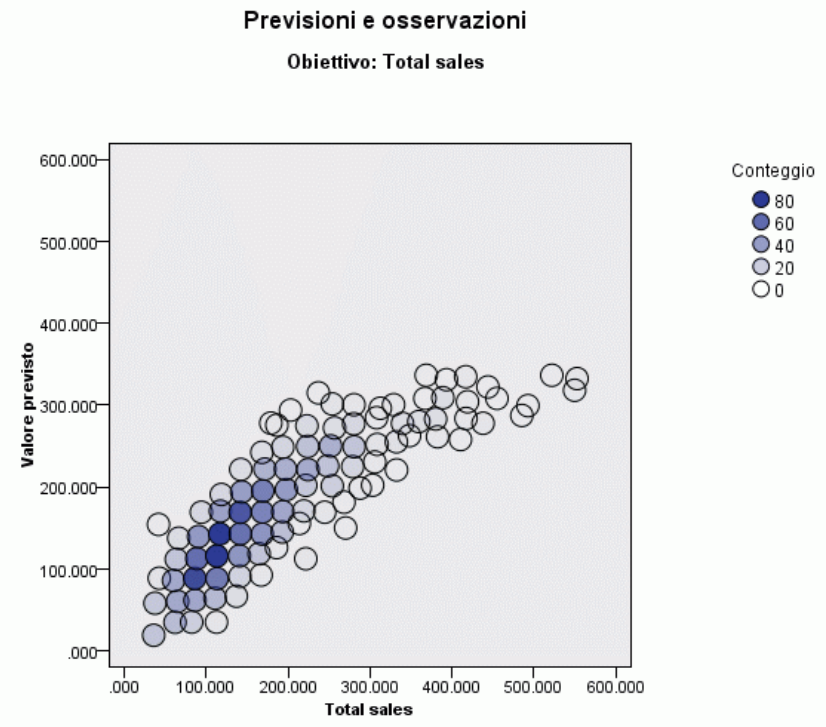
Figura 15-10
Visualizzazione Importanza predittore



Di solito è opportuno concentrare la modellazione sui campi predittori più rilevanti, lasciando perdere o ignorando i meno importanti. In questo senso può essere utile il grafico dell'importanza dei predittori, che indica l'importanza relativa di ciascun predittore nella stima del modello. Dal momento che i valori sono relativi, la somma dei valori di tutti i predittori visualizzati è pari a 1,0. L'importanza dei predittori non ha nulla a che vedere con la precisione del modello. Riguarda unicamente l'importanza di ciascun predittore per l'elaborazione di una previsione, non il grado di precisione di quest'ultima.

Previsioni e osservazioni

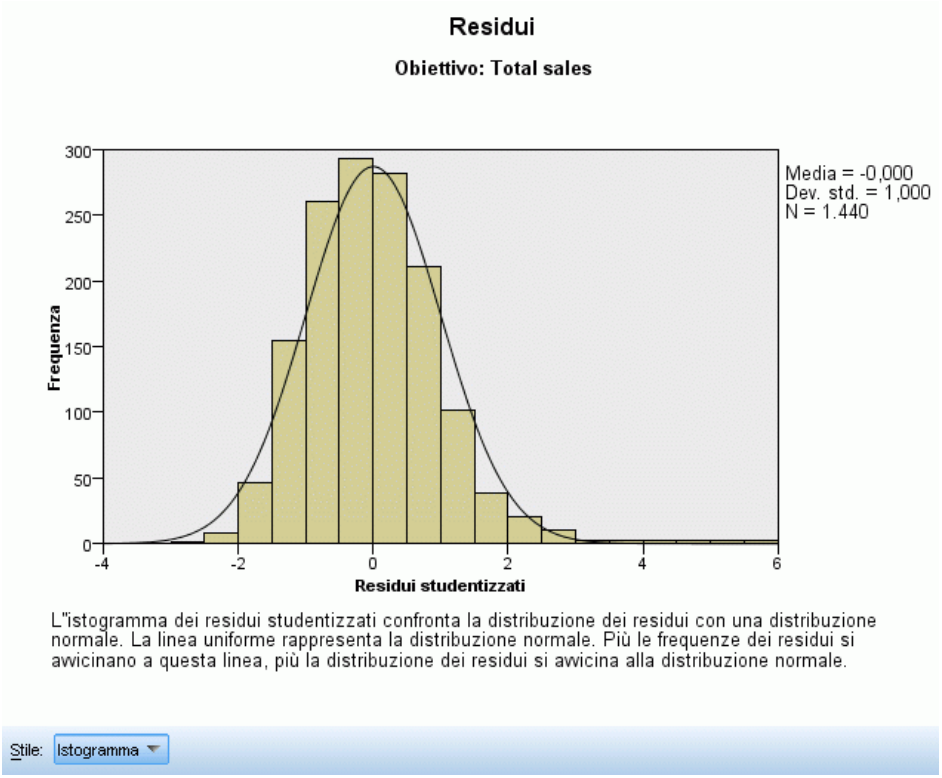
Figura 15-11
Visualizzazione Previsioni e osservazioni



Visualizza un grafico a dispersione in bin dei valori previsti sull'asse verticale in base ai valori osservati sull'asse orizzontale. Idealmente, i punti devono trovarsi su una linea a 45 gradi; questa visualizzazione segnala se vi sono record che presentano particolari problemi di previsione da parte del modello.

Residui

Figura 15-12
Visualizzazione Residui, stile istogramma



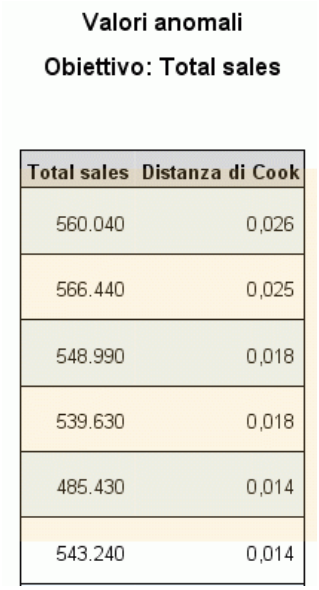
Visualizza un grafico diagnostico dei residui del modello.

Stili del grafico. Esistono diversi stili di visualizzazione, disponibili nell'elenco a discesa Stile.

- **Istogramma.** Istogramma in bin dei residui studentizzati con una sovrapposizione della distribuzione normale. I modelli lineari presumono che i residui abbiano una distribuzione normale, quindi l'istogramma idealmente dovrebbe avvicinarsi molto alla linea uniforme.
- **Grafico P-P.** Grafico probabilità-probabilità in bin che confronta i residui studentizzati con una distribuzione normale. Se la pendenza del tracciato dei punti è minore rispetto alla linea normale, i residui mostrano una maggiore variabilità rispetto a una distribuzione normale. Se la pendenza è maggiore, i residui risultano meno variabili rispetto a una distribuzione normale. Se i punti del tracciato formano una curva a S, la distribuzione dei residui è asimmetrica.

Valori anomali

Figura 15-13
Visualizzazione Valori anomali



Valori anomali

Obiettivo: Total sales

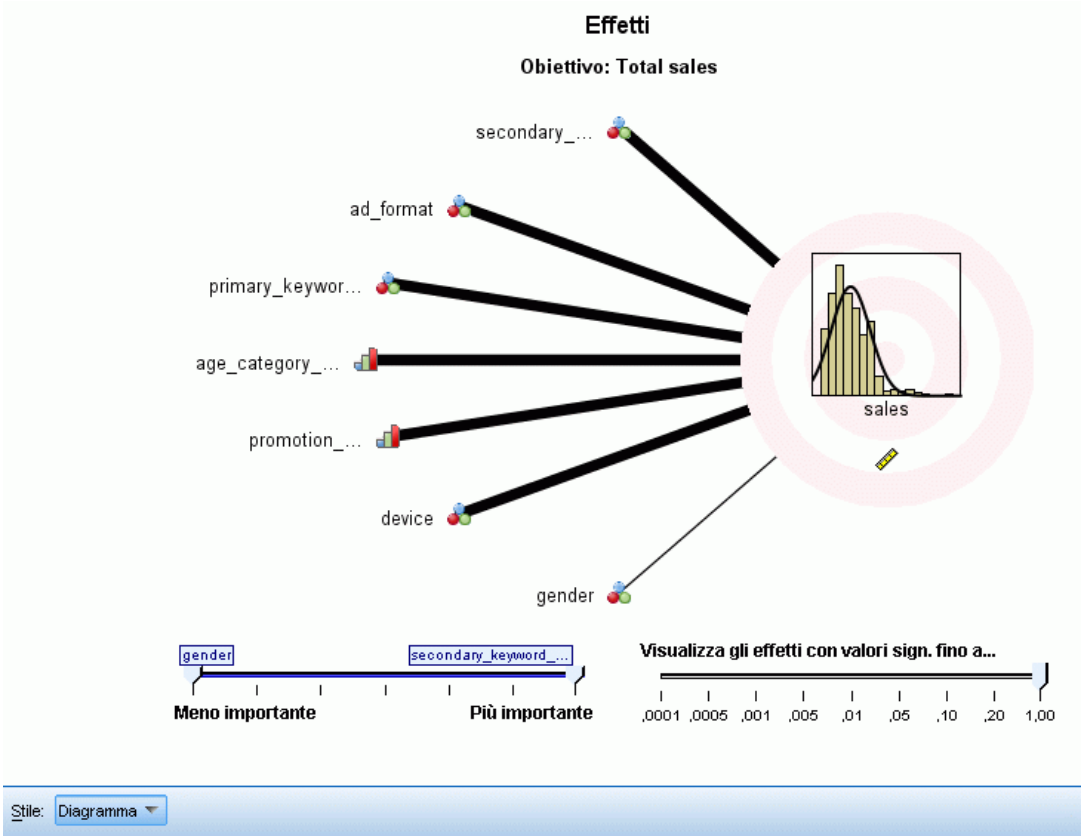
Total sales	Distanza di Cook
560.040	0,026
566.440	0,025
548.990	0,018
539.630	0,018
485.430	0,014
543.240	0,014

Questa tabella contiene i record che influenzano in modo anomalo il modello e visualizza ID dei record, se specificato nella scheda Campi, valore dell'obiettivo e distanza di Cook. La distanza di Cook è una misura di quanto cambierebbero i residui di tutti i record se un particolare record fosse escluso dal calcolo dei coefficienti del modello. Un valore elevato per la distanza di Cook indica che l'esclusione di un record modifica i coefficienti in modo sostanziale e deve quindi essere considerata come fattore influente.

I record influenti devono essere esaminati con attenzione per determinare se è possibile dare agli stessi meno peso nella stima del modello, troncare i valori anomali in corrispondenza di una soglia accettabile o eliminare completamente i record influenti.

Effetti

Figura 15-14
Visualizzazione Effetti, stile diagramma



Questa visualizzazione mostra le dimensioni di ogni effetto nel modello.

Stili. Esistono diversi stili di visualizzazione, disponibili nell'elenco a discesa Stile.

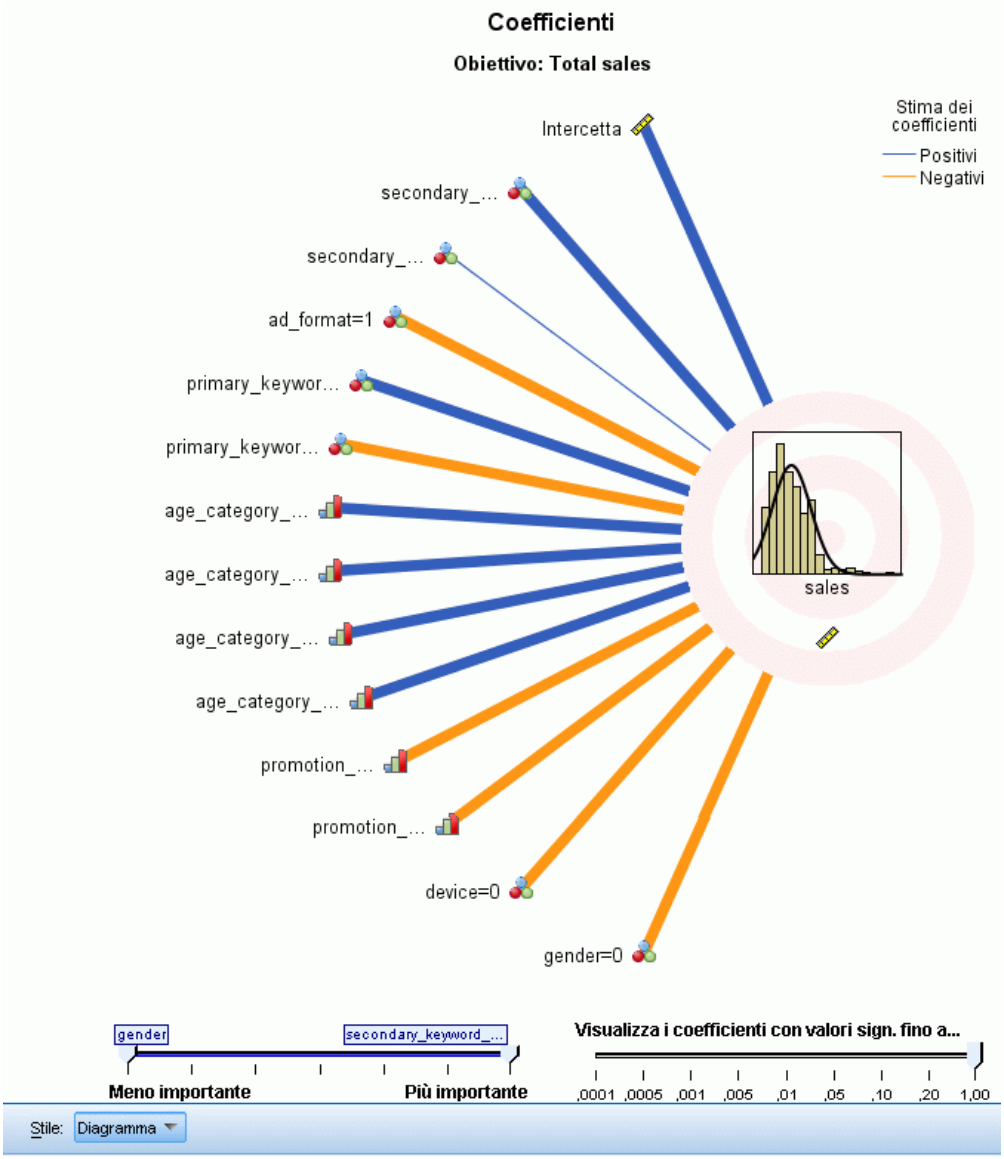
- **Diagramma.** Si tratta di un grafico in cui gli effetti vengono ordinati dall'alto verso il basso in ordine decrescente in base all'importanza dei predittori. Le linee di collegamento del diagramma vengono pesate in base alla significatività degli effetti, con la maggiore ampiezza della linea corrispondente agli effetti più significativi (minori valori *p*). Se si passa il mouse sopra una linea di collegamento viene visualizzata una descrizione che mostra il valore *p* e l'importanza dell'effetto. È l'impostazione di default.
- **Tabella.** Tabella ANOVA per gli effetti generali e specifici del modello. Gli effetti specifici sono ordinati dall'alto verso il basso in ordine decrescente in base all'importanza dei predittori. Si noti che, per impostazione predefinita, la tabella viene compressa per mostrare solo i risultati del modello generale. Per vedere i risultati dei singoli effetti del modello, fare clic sulla cella Modello corretto nella tabella.

Importanza predittore. Il dispositivo di scorrimento Importanza predittore consente di determinare quali predittori mostrare nella visualizzazione. L'azione del dispositivo di scorrimento non modifica il modello, consente semplicemente di evidenziare i predittori più importanti. Per default sono visualizzati i primi 10 effetti.

Significatività. Dispositivo di scorrimento che consente di selezionare ulteriori effetti da visualizzare, oltre a quelli selezionati in base all'importanza dei predittori. Gli effetti con valori di significatività superiori al valore del dispositivo di scorrimento vengono nascosti. L'azione del dispositivo di scorrimento non modifica il modello, consente semplicemente di evidenziare gli effetti più importanti. Il valore predefinito è 1,00, ovvero gli effetti non vengono filtrati in base alla significatività.

Coefficienti

Figura 15-15
Visualizzazione Coefficienti, stile diagramma



Questa visualizzazione mostra il valore di ogni coefficiente nel modello. Si noti che i fattori (predittori categoriali) sono codificati mediante un indicatore nel modello, in modo tale che agli **effetti** contenenti fattori possano essere associati più **coefficienti**, uno per ogni categoria esclusa la categoria corrispondente al parametro ridondante (riferimento).

Stili. Esistono diversi stili di visualizzazione, disponibili nell'elenco a discesa Stile.

- **Diagramma.** Grafico che prima visualizza l'intercetta e quindi ordina gli effetti dall'alto verso il basso in ordine decrescente in base all'importanza dei predittori. Negli effetti che contengono fattori i coefficienti vengono ordinati in ordine crescente in base ai valori dei

dati. Le linee di collegamento del diagramma sono colorate in base al segno del coefficiente (vedere la chiave del diagramma) e pesate in base alla significatività dei coefficienti, con la maggiore ampiezza della linea corrispondente ai coefficienti più significativi (minori valori *p*). Se si passa il mouse sopra una linea di collegamento viene visualizzata una descrizione che mostra il valore del coefficiente, il suo valore *p* e l'importanza dell'effetto a cui è associato il parametro. Questo è lo stile di default.

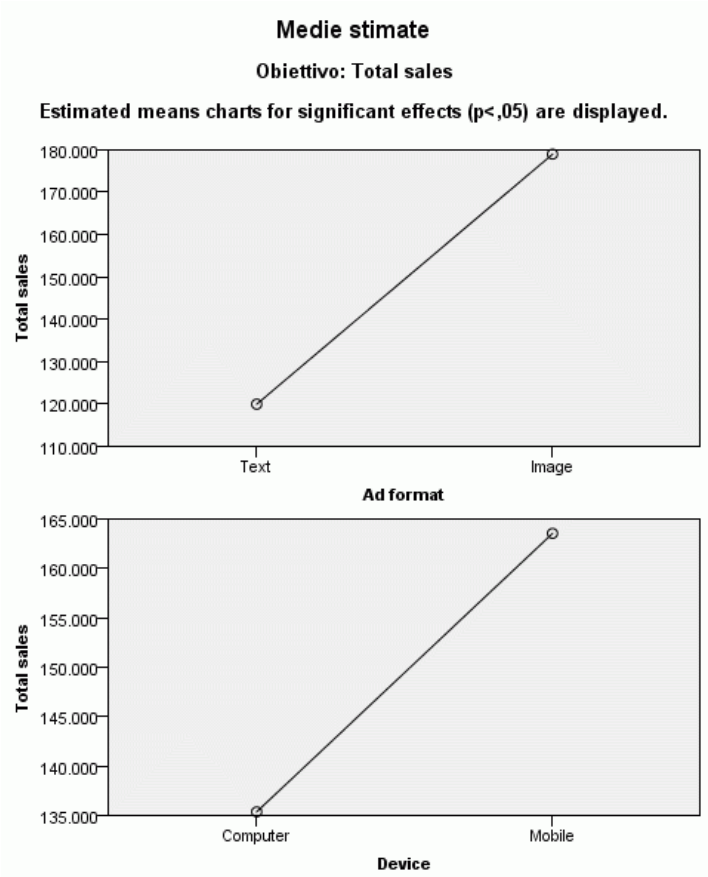
- **Tabella.** Indica i valori, i test di significatività e gli intervalli di confidenza per i singoli coefficienti del modello. Dopo l'intercetta, gli effetti vengono ordinati dall'alto verso il basso in ordine decrescente in base all'importanza dei predittori. Negli effetti che contengono fattori i coefficienti vengono ordinati in ordine crescente in base ai valori dei dati. Si noti che, per impostazione predefinita, la tabella viene compressa per mostrare solo il coefficiente, la significatività e l'importanza di ogni parametro del modello. Per vedere l'errore standard, la statistica *t* e l'intervallo di confidenza, fare clic sulla cella Coefficiente nella tabella. Se si passa il mouse sopra il nome di un parametro del modello nella tabella viene visualizzata una descrizione che mostra il nome del parametro, l'effetto a cui è associato il parametro e, per i predittori categoriali, le etichette dei valori associati al parametro del modello. Ciò può essere particolarmente utile per visualizzare le nuove categorie create quando la preparazione automatica dei dati unisce categorie simili di un predittore categoriale.

Importanza predittore. Il dispositivo di scorrimento Importanza predittore consente di determinare quali predittori mostrare nella visualizzazione. L'azione del dispositivo di scorrimento non modifica il modello, consente semplicemente di evidenziare i predittori più importanti. Per default sono visualizzati i primi 10 effetti.

Significatività. Dispositivo di scorrimento che consente di selezionare ulteriori coefficienti da visualizzare, oltre a quelli selezionati in base all'importanza dei predittori. I coefficienti con valori di significatività superiori al valore del dispositivo di scorrimento vengono nascosti. L'azione del dispositivo di scorrimento non modifica il modello, consente semplicemente di evidenziare i coefficienti più importanti. Il valore di default è 1,00, ovvero i coefficienti non vengono filtrati in base alla significatività.

Medie stimate

Figura 15-16
Visualizzazione Medie stimate



Vengono visualizzati grafici relativi ai predittori significativi. Ogni grafico visualizza il valore del modello stimato relativo all'obiettivo sull'asse verticale per ogni valore del predittore sull'asse orizzontale, mantenendo costanti tutti gli altri predittori. È una visualizzazione utile degli effetti dei coefficienti di ciascun predittore sull'obiettivo.

Nota: se non sono presenti predittori significativi, non vengono generate medie stimate.

Riepilogo di creazione dei modelli

Figura 15-17
Visualizzazione Riepilogo di creazione dei modelli, algoritmo stepwise in avanti

Riepilogo di creazione dei modelli

Obiettivo: Total sales

		Passaggio						
		1	2	3	4	5	6	7
Criterio di informazione		11.949,413	11.597,758	11.347,000	11.118,878	10.965,287	10.816,338	10.803,021
Effetto	secondary_keyword_transformed	✓	✓	✓	✓	✓	✓	✓
	ad_format		✓	✓	✓	✓	✓	✓
	primary_keyword_transformed			✓	✓	✓	✓	✓
	age_category_transformed				✓	✓	✓	✓
	promotion_transformed					✓	✓	✓
	device						✓	✓
	gender							✓

Il metodo di creazione dei modelli è stepwise in avanti utilizzando il criterio di informazione.
Un segno di spunta significa che in questo passaggio l'effetto è nel modello.

Quando si sceglie un algoritmo di selezione del modello diverso da Nessuno in Selezione modello, vengono visualizzati alcuni dettagli del processo di creazione del modello.

Stepwise in avanti. Se l'algoritmo di selezione è stepwise in avanti, la tabella visualizza gli ultimi 10 passaggi dell'algoritmo stepwise. Per ogni passaggio, vengono visualizzati il valore del criterio di selezione e gli effetti nel modello in corrispondenza di tale passaggio. Ciò consente di verificare l'entità del contributo di ogni passaggio al modello. Ciascuna colonna consente di ordinare le righe in modo tale da individuare più facilmente gli effetti presenti nel modello in un determinato passaggio.

Sottoinsiemi migliori. Se l'algoritmo di selezione è sottoinsiemi migliori, la tabella visualizza i primi 10 modelli. Per ogni modello, vengono visualizzati il valore del criterio di selezione e gli effetti presenti nel modello. Ciò consente di verificare la stabilità dei modelli più importanti. Se i modelli tendono ad avere molti effetti simili con poche differenze, il modello migliore può essere considerato affidabile, se invece i modelli contengono effetti molto diversi, alcuni effetti potrebbero essere troppo simili e sarebbe opportuno combinarli (o eliminarne uno). Ciascuna colonna consente di ordinare le righe in modo tale da individuare più facilmente gli effetti presenti nel modello in un determinato passaggio.

Capitolo
16

Regressione lineare

La regressione lineare consente di stimare i coefficienti dell'equazione lineare, includendo una o più variabili indipendenti, che prevedono al meglio il valore della variabile dipendente. Ad esempio, è possibile tentare di prevedere le vendite annuali di un rappresentante (la variabile dipendente) in base a variabili indipendenti quali l'età, gli studi e gli anni di esperienza lavorativa.

Esempio. Il numero di partite vinte da una squadra di basket in una stagione è correlato al numero medio di punti effettuati dalla squadra per partita? Un grafico a dispersione indica che queste variabili sono correlate in modo lineare. Il numero di partite vinte e il numero medio di punti effettuati dalla squadra avversaria sono anch'essi correlati in modo lineare. Queste variabili hanno una relazione negativa. Al crescere del numero delle partite vinte, diminuisce il numero medio di punti effettuati dall'avversario. Con la regressione lineare, è possibile modellare la relazione di queste variabili. È possibile utilizzare un modello valido per stimare quante partite vinceranno le squadre.

Statistiche. Per ogni variabile: numero di casi validi, media e deviazione standard. Per ogni modello: coefficienti di regressione, matrice di correlazione, correlazioni di ordine zero e parziali, *R* multipli, $R^{quadrato}$, R^2 corretto, variazioni $R^{quadrato}$, errore standard della stima, tabella di analisi della varianza, valori attesi e residui. Inoltre, intervalli di confidenza al 95% per ogni coefficiente di regressione, matrice di varianza-covarianza, fattore d'inflazione della varianza, tolleranza, test di Durbin-Watson, misure di distanza (Mahalanobis, Cook e valori di influenza), DiffBeta, DiffAdatt, intervalli di stima e diagnostiche per casi. Grafici: grafici a dispersione, grafici parziali, istogrammi e grafici di probabilità normale.

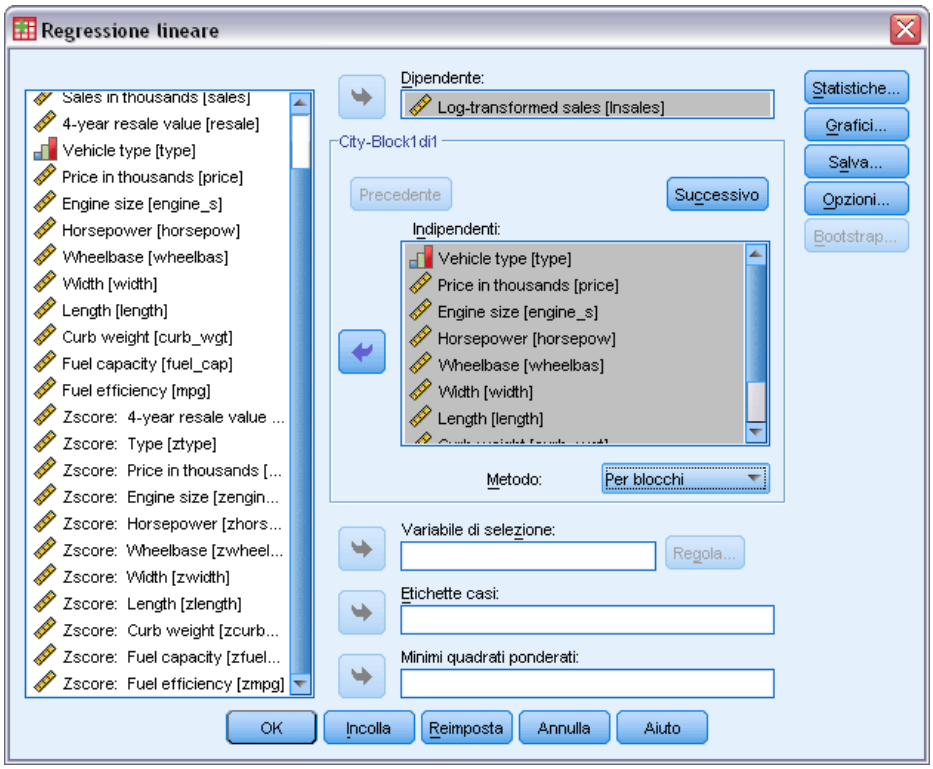
Dati. Le variabili dipendenti ed indipendenti devono essere quantitative. È necessario che le variabili categoriali, come la religione, l'età o la regione di residenza, siano ricodificate come variabili binarie (fittizie) o altri tipi di variabili di contrasto.

Assunzioni. Per ciascun valore della variabile indipendente, la distribuzione della variabile dipendente deve essere normale. La varianza della distribuzione della variabile dipendente deve essere costante per tutti i valori della variabile indipendente. La relazione tra la variabile dipendente e ogni variabile indipendente deve essere lineare e tutte le osservazioni devono essere indipendenti.

Per ottenere un'analisi della regressione lineare

- Dai menu, scegliere:
  Analizza > Regression > Lineare...

Figura 16-1
Finestra di dialogo Regressione lineare



- Nella finestra di dialogo Regressione lineare, selezionare una variabile dipendente numerica.
- Selezionare una o più variabili indipendenti numeriche.

Se lo si desidera, è possibile:

- Raggruppare le variabili indipendenti in blocchi e specificare metodi di inserimento differenti per sottogruppi di variabili diversi.
- Scegliere una variabile di selezione per limitare l'analisi a un sottoinsieme di casi con valori particolari per questa variabile.
- Selezionare una variabile di casi per l'identificazione di punti nei grafici.
- Selezionare una variabile di ponderazione per WLS per un'analisi dei minimi quadrati ponderati.

Minimi quadrati ponderati (WLS). Consente di ottenere un modello dei minimi quadrati ponderati. I dati vengono ponderati in base al reciproco della loro varianza. Le osservazioni con varianza elevata hanno un peso minore nell'analisi rispetto a quelle con varianza ridotta. Casi con valore 0, negativo o mancante per la variabile di ponderazione saranno esclusi dall'analisi.

Metodi di selezione della variabile di regressione lineare

La selezione del metodo consente di specificare come vengono inserite nell'analisi le variabili indipendenti. Utilizzando diversi metodi, è possibile creare molteplici modelli di regressione dallo stesso insieme di variabili.

- **Per blocchi (Regressione).** Una procedura per la selezione delle variabili nella quale tutte le variabili di un blocco sono inserite in un unico passo.
- **Per passi.** Ad ogni passo viene inserita la variabile indipendente non presente nell'equazione che ha la più bassa probabilità di F, se tale probabilità è sufficientemente piccola. Le variabili già presenti nell'equazione di regressione vengono rimosse se la loro probabilità di F diviene sufficientemente elevata. Il metodo termina quando nessuna variabile rispetta il criterio di inserimento o quello di rimozione.
- **Rimozione.** Una procedura per la selezione di variabili in cui tutte le variabili di un blocco sono rimosse in un solo passo.
- **Eliminazione all'indietro.** Una procedura di selezione di variabili nella quale tutte le variabili vengono inserite nell'equazione e poi rimosse sequenzialmente. La variabile con la più bassa correlazione parziale rispetto alla variabile dipendente viene considerata la prima da rimuovere e viene rimossa se soddisfa il criterio di eliminazione. Dopo la rimozione della prima variabile, la variabile con la più bassa correlazione parziale tra quelle rimaste nell'equazione viene considerata come la prossima da eliminare. La procedura termina quando nell'equazione nessuna variabile soddisfa il criterio di rimozione.
- **Selezione in avanti.** Una procedura di selezione delle variabili nella quale le variabili vengono inserite in modo sequenziale all'interno del modello. La prima variabile da inserire nell'equazione è quella con la più elevata correlazione positiva o negativa con la variabile dipendente. Questa variabile viene inserita nell'equazione solo se soddisfa il criterio di inserimento. Se è stata inserita la prima variabile, viene considerata come successiva la variabile indipendente non presente nell'equazione che ha la più elevata correlazione parziale. La procedura termina quando non ci sono più variabili che soddisfano il criterio di inserimento.

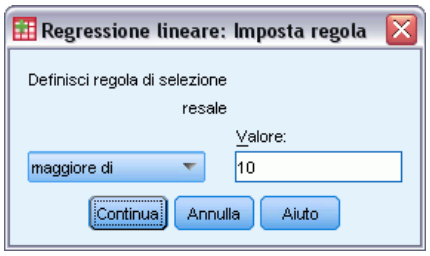
I valori di significatività dell'output si basano sull'adattamento di un singolo modello. Pertanto, i valori di significatività in genere non sono validi quando viene utilizzato un metodo stepwise (per passi, avanti o indietro).

Tutte le variabili devono soddisfare il criterio di tolleranza per essere inserite nell'equazione, indipendentemente dal metodo di inserimento specificato. Il livello di tolleranza predefinito è 0,0001. Una variabile non viene inserita se può far sì che la tolleranza di un'altra variabile già nel modello non rientri nel criterio di tolleranza già stabilito.

Tutte le variabili indipendenti selezionate vengono aggiunte a un solo modello di regressione. È tuttavia possibile specificare diversi metodi di inserimento per diversi sottoinsiemi di variabili. Ad esempio, è possibile inserire un blocco di variabili nel modello di regressione utilizzando la selezione per passi e un secondo blocco utilizzando la selezione in avanti. Per aggiungere un secondo blocco di variabili a un modello di regressione, fare clic su Avanti.

Regressione lineare: Imposta regola

Figura 16-2
Finestra di dialogo Regressione lineare: Imposta regola



Nell'analisi verranno inseriti i casi definiti dalla regola di selezione impostata. Ad esempio, se si seleziona la variabile Uguale a e si immette 5 per il valore, nell'analisi verranno inclusi solo i casi in cui la variabile selezionata ha un valore uguale a 5. È inoltre possibile specificare un valore stringa.

Regressione lineare: grafici

Figura 16-3
Finestra di dialogo Regressione lineare: grafici

I grafici possono facilitare la validazione delle ipotesi di normalità, linearità e uguaglianza delle varianze. I grafici sono inoltre utili per la rilevazione di valori anomali, osservazioni insolite e casi di influenza. Dopo averli salvati come nuove variabili, sarà possibile utilizzare valori attesi, residui e altre diagnostiche disponibili nell'Editor dei dati per la creazione di grafici con le variabili indipendenti. Sono disponibili i seguenti grafici:

Grafici a dispersione. Consentono di rappresentare due elementi qualsiasi tra i seguenti: la variabile dipendente, i valori attesi standardizzati, i residui standardizzati, i residui cancellati, i valori attesi corretti, i residui studentizzati o i residui cancellati studentizzati. Rappresentare nel grafico i residui standardizzati e i valori attesi standardizzati per verificare la linearità e l'uguaglianza delle varianze.

Elenco di variabili sorgente. Elenca la variabile dipendente (DEPENDNT) e le seguenti variabili (valori stimati e residui): valori stimati standardizzati (*ZPRED), residui standardizzati (*ZRESID), residui cancellati (*DRESID), valori stimati corretti (*ADJPRED), residui studentizzati (*SRESID), residui cancellati studentizzati (*SDRESID).

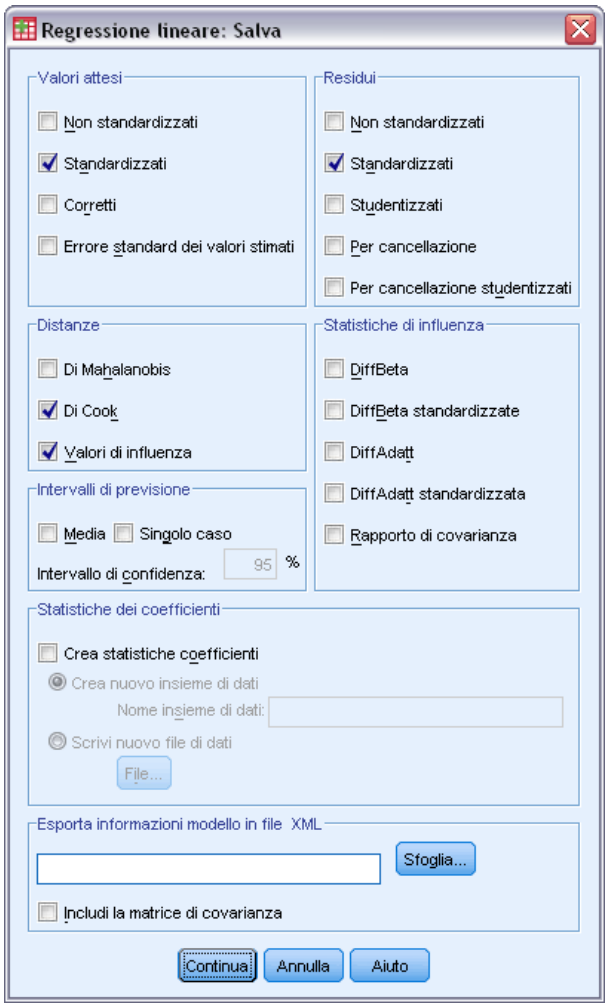
Produci tutti i grafici parziali. Consente di visualizzare i grafici a dispersione dei residui di ogni variabile indipendente e i residui della variabile dipendente quando entrambe le variabili sono regresse separatamente dal resto delle variabili indipendenti. Per la creazione di un grafico parziale è necessario che nell'equazione siano rappresentate almeno due variabili indipendenti.

Grafici dei residui standardizzati . Consentono di ottenere istogrammi dei residui standardizzati e grafici di probabilità normale confrontando la distribuzione dei residui standardizzati e la distribuzione normale.

Se vengono richiesti grafici, verranno visualizzate statistiche riassuntive per i valori attesi standardizzati e per i residui standardizzati (*ZPRED e *ZRESID).

Regressione lineare: Per salvare nuove variabili

Figura 16-4
Finestra di dialogo Regressione lineare: Salva



È possibile salvare valori attesi, residui e altre statistiche utili per la diagnostica. Ogni selezione aggiunge una o più nuove variabili al file dati attivo.

Valori attesi. I valori previsti dal modello di regressione per ogni caso.

- **Non standardizzati.** I valori risultanti dal modello per la variabile dipendente e per ciascun caso.
- **Standardizzati.** Una trasformazione di ciascun valore atteso nella sua forma standardizzata. ovvero il valore atteso medio viene sottratto dal valore atteso e la differenza viene divisa per la deviazione standard dei valori attesi. I valori attesi standardizzati hanno media 0 e deviazione standard 1.

- **Corretti.** Il valore atteso per un caso quando quel caso è escluso dal calcolo dei coefficienti di regressione.
- **Errore standard dei valori stimati.** L'errore standard della media per i valori stimati. Una stima della deviazione standard del valore medio della variabile dipendente per i casi che hanno gli stessi valori delle variabili indipendenti.

Distanze. Misure per l'identificazione dei casi con combinazioni di valori insolite per le variabili indipendenti e dei casi che possono avere un notevole peso sul modello di regressione.

- **Di Mahalanobis.** Una misura della distanza di un caso dalla media di tutti i casi per le variabili indipendenti. Un'elevata distanza di Mahalanobis indica che un caso include valori estremi per una o più variabili indipendenti.
- **Di Cook.** Una misura di quanto cambierebbero i residui di tutti i casi se un particolare caso fosse escluso dal calcolo dei coefficienti di regressione. Valori alti indicano che l'esclusione di un caso dal calcolo dei coefficienti di regressione ne modificherebbe sostanzialmente il valore.
- **Valori di influenza.** Una misura dell'influenza di un dato sull'adattamento della regressione. L'influenza centrata varia da 0 (nessuna influenza sull'adattamento) a (N-1)/N.

Intervalli di previsione. I limiti superiore ed inferiore per gli intervalli di previsione singoli e medi.

- **Media.** Limiti inferiore e superiore (due variabili) per l'intervallo di stima della risposta media prevista.
- **Singolo caso.** Limiti inferiore e superiore (due variabili) per l'intervallo di stima della variabile dipendente per un singolo caso.
- **Intervallo di confidenza.** Immettere un valore compreso fra 1 e 99,99 per specificare l'intervallo di confidenza per i due intervalli di stima. Per disporre di questa opzione è necessario aver selezionato Media o Individuale. I valori tipici dell'intervallo di confidenza sono 90, 95 e 99.

Residui. Il valore effettivo della variabile dipendente meno il valore atteso dall'equazione di regressione.

- **Non standardizzati.** La differenza tra un valore osservato e il valore stimato dal modello.
- **Standardizzati.** Il residuo diviso per una stima della deviazione standard. Il residuo standardizzato, conosciuto anche come residuo di Pearson, ha media 0 e deviazione standard 1.
- **Studentizzati.** Il residuo diviso per una stima della sua deviazione standard che varia da caso a caso, a seconda della distanza tra i valori assunti per questo caso dalle variabili indipendenti e le medie delle variabili indipendenti.
- **Per cancellazione.** Il residuo per un caso se quel caso venisse escluso dal calcolo dei coefficienti di regressione. È la differenza tra il valore della variabile dipendente e il valore stimato corretto.
- **Per cancellazione studentizzati.** Il residuo cancellato per un caso diviso per il suo errore standard. La differenza tra un residuo cancellato studentizzato e il suo corrispondente residuo studentizzato indica quanta differenza produca l'eliminazione di un caso sulla stima del medesimo.

Statistiche di influenza. La modifica nei coefficienti di regressione (DiffBeta) e nei valori attesi (DiffAdatt) che risultano dall'esclusione di un particolare caso. I valori DiffBeta e DiffAdatt standardizzati sono anche disponibili con il rapporto di covarianza.

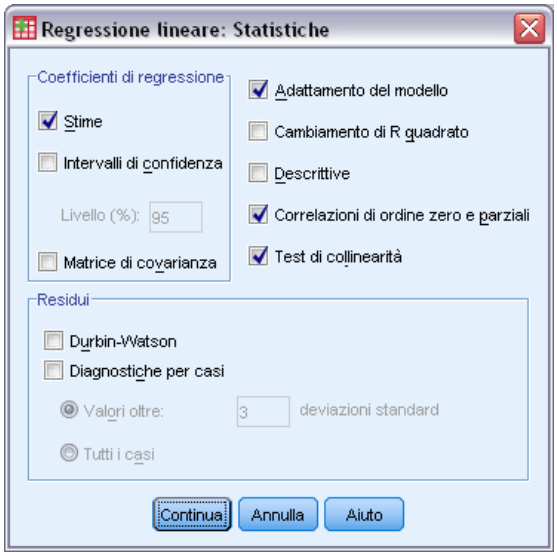
- **Differenza in beta.** Variazione del coefficiente di regressione quando un caso particolare viene eliminato dall'analisi. Viene calcolato un valore per ogni termine del modello, incluso il termine costante.
- **DiffBeta standardizzate.** La differenza standardizzata nel valore beta. La variazione di un coefficiente di regressione quando un caso viene rimosso dall'analisi. Possono essere esaminati i casi con valore assoluto superiore a 2 diviso per la radice quadrata di N, dove N è il numero dei casi. Viene calcolato un valore per ogni termine del modello, incluso il termine costante.
- **DiffAdatt.** La differenza nel valore adattato è il cambiamento del valore previsto quando un caso viene escluso dall'analisi.
- **DiffAdatt standardizzata.** La differenza standardizzata nel valore adattato. La variazione del valore stimato quando un caso viene rimosso dall'analisi. Si possono esaminare valori standardizzati maggiori in valore assoluto a 2 per la radice quadrata di p/N, dove p è il numero di parametri del modello e N è il numero di casi.
- **Rapporto di covarianza.** Rapporto fra il determinante della matrice di covarianza con un caso escluso dal calcolo dei coefficienti di regressione ed il determinante con tutti i casi inclusi. Se il rapporto è vicino a 1, il caso non altera significativamente la matrice di covarianza.

Statistiche dei coefficienti. Salva i coefficienti di regressione in un file di dati. I file di dati possono anche essere riutilizzati nella stessa sessione, ma non vengono salvati come file a meno che siano stati salvati come tali alla fine della sessione. I nomi degli insiemi di dati devono essere conformi alle regole dei nomi delle variabili.

Esporta informazioni modello in file XML. Le stime dei parametri e, se si desidera, le relative covarianze vengono esportati nel file specificato in formato XML (PMML). È possibile utilizzare questo file di modello per applicare le informazioni del modello ad altri file di dati per il calcolo del punteggio.

Regressione lineare: Statistiche

Figura 16-5
Finestra di dialogo Statistiche



Sono disponibili le seguenti statistiche:

Coefficienti di regressione. Stime consente di visualizzare il coefficiente di regressione *B*, l'errore standard di *B*, il coefficiente beta standardizzato, il valore *t* per *B*e il livello di significatività a due code di *t*. Intervalli di confidenza visualizza gli intervalli di confidenza con il livello di confidenza specificato per ciascun coefficiente di regressione o matrice di covarianza. Matrice di covarianza consente di visualizzare una matrice di varianza-covarianza dei coefficienti di regressione con le covarianze esterne alla diagonale e le varianze sulla diagonale. Viene inoltre visualizzata una matrice di correlazione.

Adattamento del modello. Vengono elencate le variabili inserite ed eliminate dal modello e vengono visualizzate le seguenti statistiche di bontà dell'adattamento: *R* multipli, $R^{quadrato}$ e $R^{quadrato}$ corretto, errore standard della stima e una tabella di analisi della varianza.

Cambiamento di R quadrato. La variazione nella statistica $R^{quadrato}$ che viene prodotta aggiungendo o eliminando una variabile indipendente. Se la variazione di $R^{quadrato}$ associata ad una variabile è elevata, ciò significa che la variabile è un valido stimolatore della variabile dipendente.

Descrittive. Fornisce il numero di casi validi, la media e la deviazione standard per ogni variabile nell'analisi. Vengono inoltre visualizzati una matrice di correlazione con un livello di significatività a una coda e il numero di casi per ogni correlazione.

Correlazione parziale. Ciò che rimane della correlazione fra due variabili, dopo aver rimosso gli effetti della loro reciproca correlazione con altre variabili. Ad esempio, la correlazione fra la variabile dipendente e una variabile indipendente dopo aver rimosso da entrambe gli effetti lineari delle altre variabili indipendenti incluse nel modello.

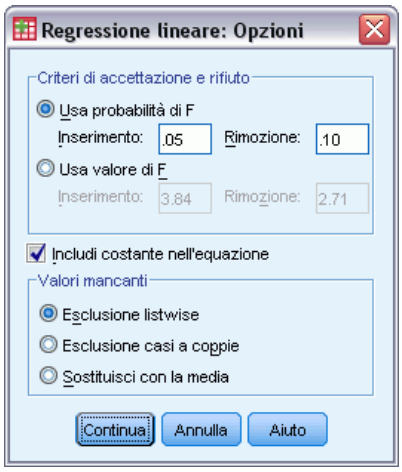
Correlazione parziale. La correlazione fra la variabile dipendente e una variabile indipendente, dopo aver rimosso dalla variabile indipendente gli effetti lineari della correlazione con altre variabili indipendenti. Correlata alla variazione di R-quadrato quando una variabile viene aggiunta a un'equazione. A volte detta correlazione semiparziale.

Diagnostiche di collinearità. La collinearità (o multicollinearità) è la situazione in cui una delle variabili indipendenti è una funzione lineare di altre variabili indipendenti. Consente di visualizzare gli autovalori della matrice dei prodotti incrociati scalata e non centrata, l'indice di collinearità e le proporzioni della decomposizione della varianza con i fattori di inflazione della varianza (VIF) e le tolleranze per le singole variabili.

Residui. Consente di visualizzare il test di Durbin-Watson per la correlazione seriale dei residui e la diagnostica per i casi che corrispondono al criterio di selezione (valori anomali superiori a *n* deviazioni standard).

Regressione lineare: Opzioni

Figura 16-6
Finestra di dialogo Regressione lineare: Opzioni



Sono disponibili le seguenti opzioni:

Criteri di accettazione e rifiuto. Queste opzioni vengono utilizzate quando si specifica il metodo di selezione delle variabili in avanti, indietro o per passi. È possibile inserire o eliminare le variabili dal modello in base alla significatività (probabilità) del valore *F* o allo stesso valore *F*.

- **Usa probabilità di F.** La variabile viene inserita nel modello se il livello di significatività del relativo valore di F è minore di quello di inserimento. La variabile viene altresì rimossa se il livello di significatività è maggiore di quello di rimozione. I valori di inserimento e di rimozione devono essere entrambi positivi e Inserimento deve essere minore di Rimozione. Alzando il valore di inserimento e/o abbassando quello di rimozione si allentano i vincoli di inclusione delle variabili nel modello.
- **Usa valore di F.** La variabile viene inserita nel modello se il relativo valore F è maggiore di quello di inserimento. La variabile viene altresì rimossa se il relativo valore F è minore di quello di rimozione. I valori di inserimento e di rimozione devono essere entrambi positivi

e Inserimento deve essere maggiore di Rimozione Abbassando il valore di inserimento e/o alzando quello di rimozione si allentano i vincoli di inclusione delle variabili nel modello.

Includi termine costante nell'equazione. Per impostazione predefinita, il modello di regressione include un termine costante. Se l'opzione è deselezionata, viene forzato il passaggio della curva di regressione per l'origine, il che avviene raramente. Alcuni risultati di una curva di regressione che passa per l'origine non sono confrontabili con i risultati della regressione che include una costante. Ad esempio R^{quadrato} non può essere interpretato nel modo usuale.

Valori mancanti. È possibile scegliere tra le opzioni seguenti:

- **Esclusione listwise.** Sono inclusi nell'analisi solo i casi con valori validi per tutte le variabili.
- **Esclusione pairwise.** Per calcolare il coefficiente di correlazione su cui si basa l'analisi della regressione vengono utilizzati i casi con dati completi per la coppia di variabili correlate. I gradi di libertà sono basati su *N* pairwise minimo.
- **Sostituisci con la media.** Per i calcoli vengono utilizzati tutti i casi e la media della variabile viene sostituita alle osservazioni mancanti.

Opzioni aggiuntive del comando REGRESSION

Il linguaggio della sintassi dei comandi consente inoltre di:

- Scrivere una matrice di correlazione o leggere una matrice anziché i dati grezzi per ottenere l'analisi di regressione (con il sottocomando `MATRIX`).
- Specificare i livelli di tolleranza (tramite il sottocomando `CRITERIA`).
- Ottenere più modelli per variabili dipendenti uguali o diverse (con i sottocomandi `METHOD` e `DEPENDENT`).
- Ottenere statistiche aggiuntive (con i sottocomandi `DESCRIPTIVES` e `STATISTICS`).

Vedere *Command Syntax Reference* per informazioni dettagliate sulla sintassi.

Capitolo
17

Regressione ordinale

La procedura Regressione ordinale consente di definire la dipendenza di una risposta ordinale politomica in un insieme di stimatori, che possono essere fattori o covariate. La struttura della regressione ordinale è basata sulla metodologia di McCullagh (1980, 1998) e la procedura è denominata `PLUM` nella sintassi.

L'analisi della regressione lineare standard comporta la riduzione al minimo della somma dei quadrati delle differenze tra una variabile di risposta (dipendente) e una combinazione ponderata di variabili stimatore (indipendenti). I coefficienti stimati riflettono il modo in cui le modifiche dei predittori influiscono sulla risposta. Si presume che la risposta sia numerica e ciò significa che le modifiche al livello di risposta sono uguali nell'intervallo della risposta. Ad esempio, la differenza di altezza tra una persona alta 150 cm e una persona alta 140 cm è di 10 cm ed equivale alla differenza di altezza tra una persona alta 210 cm e una persona alta 200 cm. Queste relazioni non sono necessariamente valide per le variabili ordinali, nelle quali la scelta e il numero delle categorie di risposta possono essere arbitrarie.

Esempio. È possibile utilizzare la regressione ordinale per studiare la reazione dei pazienti a un dosaggio medicinale. Le possibili reazioni possono essere classificate come *nessuna*, *lieve*, *moderata* o *grave*. La differenza tra una reazione lieve e una moderata è molto difficile o impossibile da quantificare ed è basata sulla percezione. In generale, la differenza tra una risposta lieve e una moderata può essere maggiore o minore della differenza tra una risposta moderata e una grave.

Statistiche e grafici. Frequenze osservate e attese e frequenze cumulate, residui di Pearson per le frequenze e frequenze cumulate, probabilità osservate e attese, probabilità osservate e probabilità cumulate attese per ciascuna categoria di risposta in base al modello covariato, correlazione asintotica e matrici di covarianza delle stime dei parametri, chi-quadrato di Pearson e chi-quadrato del rapporto di verosimiglianza, statistiche sulla bontà di adattamento, cronologia delle iterazioni, test dell'ipotesi di linee parallele, stime dei parametri, errori standard, intervalli di confidenza e statistiche R^2 di Cox e Snell, di Nagelkerke e di McFadden.

Dati. Si presume che la variabile dipendente sia ordinale e può essere numerica o stringa. L'ordinamento è determinato dalla disposizione dei valori della variabile dipendente in ordine crescente, in cui il valore più basso definisce la prima categoria. Si presume che le variabili fattore siano categoriali, mentre le covariate devono essere numeriche. Si noti che l'utilizzo di più covariate continue in genere comporta la creazione di una tabella di probabilità di cella di dimensioni molto grandi.

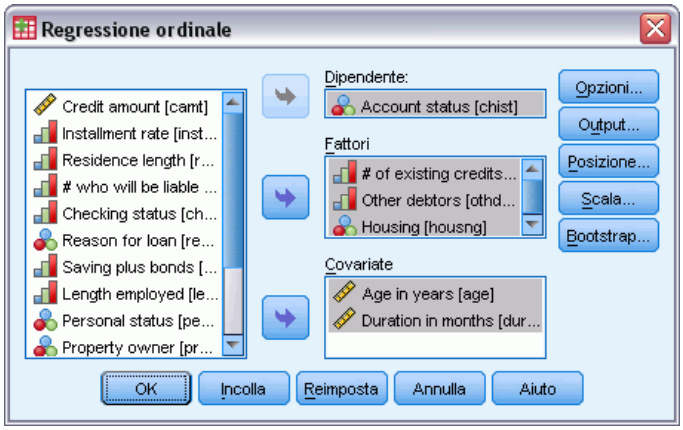
Assunzioni. È consentita una sola variabile di risposta, che deve essere specificata. Inoltre, per ciascun modello distinto di valori nelle variabili indipendenti, si presume che le risposte siano variabili multinomiali indipendenti.

Procedure correlate. La regressione logistica nominale utilizza modelli simili per le variabili dipendenti nominali.

Per ottenere una regressione ordinale

▶ Dai menu, scegliere:
Analizza > Regressione > Ordinale...

Figura 17-1
Finestra di dialogo Regressione ordinale



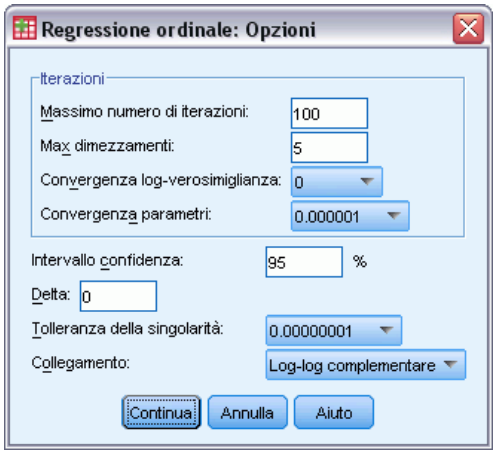
▶ Selezionare una variabile dipendente.

▶ Fare clic su OK.

Regressione ordinale: Opzioni

Nella finestra di dialogo Opzioni è possibile correggere i parametri utilizzati nell'algoritmo di stima iterativo, scegliere un livello di confidenza per le stime dei parametri e selezionare una funzione di collegamento.

Figura 17-2
Finestra di dialogo Regressione ordinale: Opzioni



Iterazioni. È possibile personalizzare l'algoritmo iterativo.

- **Max iterazioni.** Specificare un intero non negativo. Se si specifica 0, la procedura restituisce le stime iniziali.
- **Massimo numero di dimezzamenti.** Specificare un intero positivo.
- **Convergenza verosimiglianza.** L'algoritmo si interrompe se il cambiamento assoluto o relativo della verosimiglianza è inferiore a questo valore. Il criterio non viene utilizzato se si specifica 0.
- **Convergenza parametri.** L'algoritmo si interrompe se il cambiamento assoluto o relativo in ciascuna delle stime dei parametri è inferiore a questo valore. Il criterio non viene utilizzato se si specifica 0.

Intervallo di confidenza. Specificare un valore maggiore o uguale a 0 e minore di 100.

Delta. Valore aggiunto alle frequenze zero di cella. Specificare un valore non negativo inferiore a 1.

Tolleranza della singolarità Utilizzata per controllare gli stimatori a dipendenza elevata. Selezionare un valore dall'elenco delle opzioni.

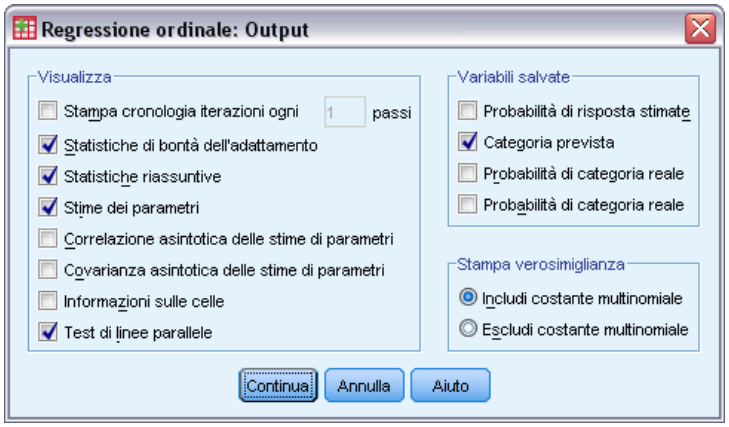
Funzione di collegamento. La funzione Collegamento è la trasformazione delle probabilità cumulate che permette di stimare il modello. Sono disponibili le cinque funzioni di collegamento descritte nella tabella che segue.

Funzione	Formato	Applicazione tipica
Logit	$\log(\xi / (1-\xi))$	Categorie distribuite uniformemente
Log-log complementare	$\log(-\log(1-\xi))$	Categorie alte più probabili
Log-log negativo	$-\log(-\log(\xi))$	Categorie basse più probabili
Probit	$\Phi^{-1}(\xi)$	La variabile latente è normalmente distribuita
Cauchit (funzione Cauchy inversa)	$\tan(\pi(\xi-0.5))$	La variabile latente ha molti valori estremi

Regressione ordinale: Output

Nella finestra di dialogo Output è possibile creare tabelle da visualizzare nel Viewer e salvare variabili nel file di lavoro.

Figura 17-3
Finestra di dialogo Regressione ordinale: Output



Visualizzazione. Consente di creare tabelle per:

- **Stampa cronologia iteraz. ogni n passi.** Stampa le stime della verosimiglianza e dei parametri per la frequenza di iterazioni di stampa specificata. La prima e l'ultima iterazione vengono sempre stampate.
- **Statistiche sulla bontà dell'adattamento.** Statistiche chi-quadrato di Pearson e del rapporto di verosimiglianza. Vengono elaborate in base alla classificazione specificata nell'elenco di variabili.
- **Statistiche riassuntive.** Statistiche R^2 di Cox e Snell, di Nagelkerke e di McFadden.
- **Stime dei parametri.** Stime dei parametri, errori standard e intervalli di confidenza.
- **Correlazione asintotica delle stime di parametri.** Matrice delle correlazioni delle stime dei parametri.
- **Covarianza asintotica delle stime di parametri.** Matrice delle covarianze delle stime dei parametri.
- **Informazioni sulle celle.** Frequenze osservate e attese e frequenze cumulate, residui di Pearson per le frequenze e le frequenze cumulate, probabilità osservate e attese, probabilità cumulate osservate e attese per ciascuna categoria di risposta in base al modello covariato. Si noti che per i modelli che includono più modelli covariati, ad esempio i modelli con covariate continue, questa opzione può creare una tabella di dimensioni molto grandi e difficile da gestire.
- **Test di linee parallele.** Test di verifica dell'ipotesi di equivalenza dei parametri di posizione nei livelli della variabile dipendente. È disponibile per il modello di sola posizione.

Variabili salvate. Salva le seguenti variabili nel file di lavoro:

- **Probabilità di risposta stimate.** Probabilità stimate dal modello per la classificazione di un modello fattore o covariato nelle categorie di risposta. Il numero di probabilità corrisponde al numero di categorie di risposta.
- **Categoria prevista.** La categoria di risposta che ha la maggiore probabilità stimata per un modello fattore o covariato.

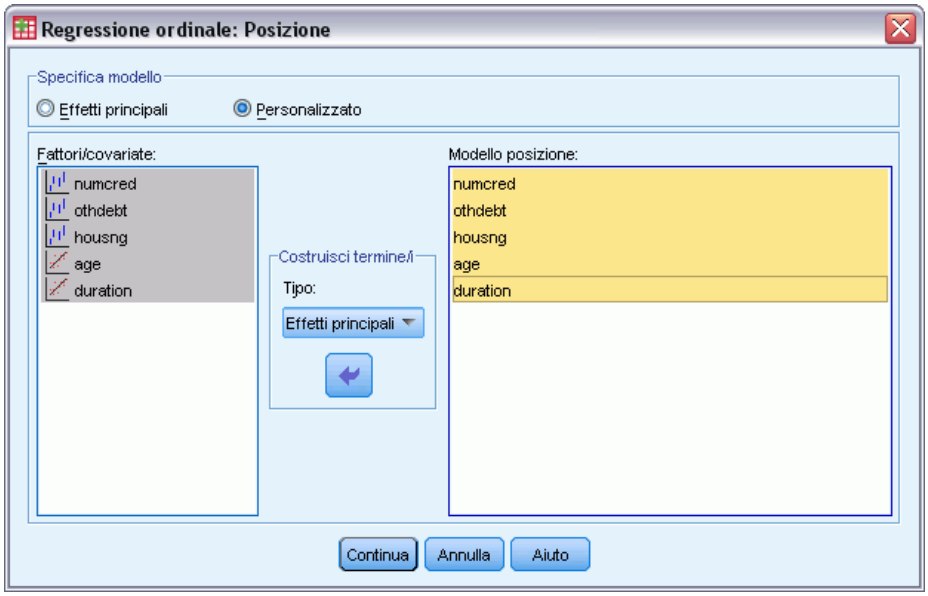
- **Probabilità di categoria prevista.** Probabilità stimata di classificazione di un modello fattore o covariato nella categoria prevista. La probabilità corrisponde inoltre al massimo di probabilità stimate del modello fattore o covariato.
- **Probabilità di categoria reale.** Probabilità stimata di classificazione di un modello fattore o covariato nella categoria reale.

Stampa verosimiglianza Controlla la visualizzazione della verosimiglianza. L'inclusione della costante multinomiale consente di ottenere il valore completo della verosimiglianza. Per confrontare i risultati dei prodotti che non includono la costante, si può scegliere di escluderla.

Regressione ordinale: Posizione

Nella finestra di dialogo Posizione è possibile specificare il modello di posizione per l'analisi.

Figura 17-4
Finestra di dialogo Regressione ordinale: Posizione



Specifica modello. Un modello a effetti principali include gli effetti principali di covariate e fattori, ma non gli effetti di interazione. È possibile creare un modello personalizzato per specificare i sottoinsiemi di interazioni di fattori o di interazioni di covariate.

Fattori/covariate. I fattori e le covariate sono elencati.

Modello posizione. Il modello dipende dagli effetti principali e di interazione selezionati.

Costruisci termini

Per i fattori e le covariate selezionati:

Interazione. Consente di creare il termine di interazione di livello maggiore rispetto a tutte le variabili selezionate. È l'impostazione predefinita.

Effetti principali. Consente di creare un termine di effetti principali per ciascuna variabile selezionata.

Tutti 2-vie. Consente di creare tutte le possibili interazioni a due vie delle variabili selezionate.

Tutti 3-vie. Consente di creare tutte le possibili interazioni a tre vie delle variabili selezionate.

Tutti 4-vie. Consente di creare tutte le possibili interazioni a quattro vie delle variabili selezionate.

Tutti 5-vie. Consente di creare tutte le possibili interazioni a cinque vie delle variabili selezionate.

Regressione ordinale: Scala

Nella finestra di dialogo Scala è possibile specificare il modello di scala per l'analisi.

Figura 17-5
Finestra di dialogo Regressione ordinale: Scala

Fattori/covariate. I fattori e le covariate sono elencati.

Modello scala. Il modello dipende dagli effetti principali e di interazione selezionati.

Costruisci termini

Per i fattori e le covariate selezionati:

Interazione. Consente di creare il termine di interazione di livello maggiore rispetto a tutte le variabili selezionate. È l'impostazione predefinita.

Effetti principali. Consente di creare un termine di effetti principali per ciascuna variabile selezionata.

Tutti 2-vie. Consente di creare tutte le possibili interazioni a due vie delle variabili selezionate.

Tutti 3-vie. Consente di creare tutte le possibili interazioni a tre vie delle variabili selezionate.

Tutti 4-vie. Consente di creare tutte le possibili interazioni a quattro vie delle variabili selezionate.

Tutti 5-vie. Consente di creare tutte le possibili interazioni a cinque vie delle variabili selezionate.

Opzioni aggiuntive del comando PLUM

Per personalizzare la procedura Regressione ordinale è possibile incollare le impostazioni selezionate in una finestra di sintassi e quindi modificare la sintassi del comando `PLUM` così ottenuta. Il linguaggio della sintassi dei comandi consente inoltre di:

- Creare test di ipotesi personalizzati specificando ipotesi nulle come combinazioni lineari dei parametri.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
18

Stima di curve

La procedura Stima di curve produce le statistiche di regressione per la stima di curve e i grafici correlati per 11 diversi modelli di regressione per la stima di curve. Per ciascuna variabile dipendente verrà creato un modello distinto. È inoltre possibile salvare come nuove variabili i valori attesi, i residui e gli intervalli di stima.

Esempio. Un provider di servizi Internet deve tener traccia della percentuale di traffico e-mail infettato da virus sulla propria rete nell'arco di un periodo di tempo specifico. Il grafico di dispersione indica che la relazione non è lineare. È necessario adattare ai dati un modello quadratico o cubico e controllare la validità delle assunzioni e la bontà di adattamento del modello.

Statistiche. Per ciascun modello: coefficienti di regressione, *R* multiplo, $R^{quadrato}$, $R^{quadrato}$ corretto, errore standard della stima, tabella di analisi della varianza, valori attesi, residui ed intervalli di stima. Modelli: lineare, logaritmico, inverso, quadratico, cubico, di potenza, composto, curva S, logistico, di crescita ed esponenziale.

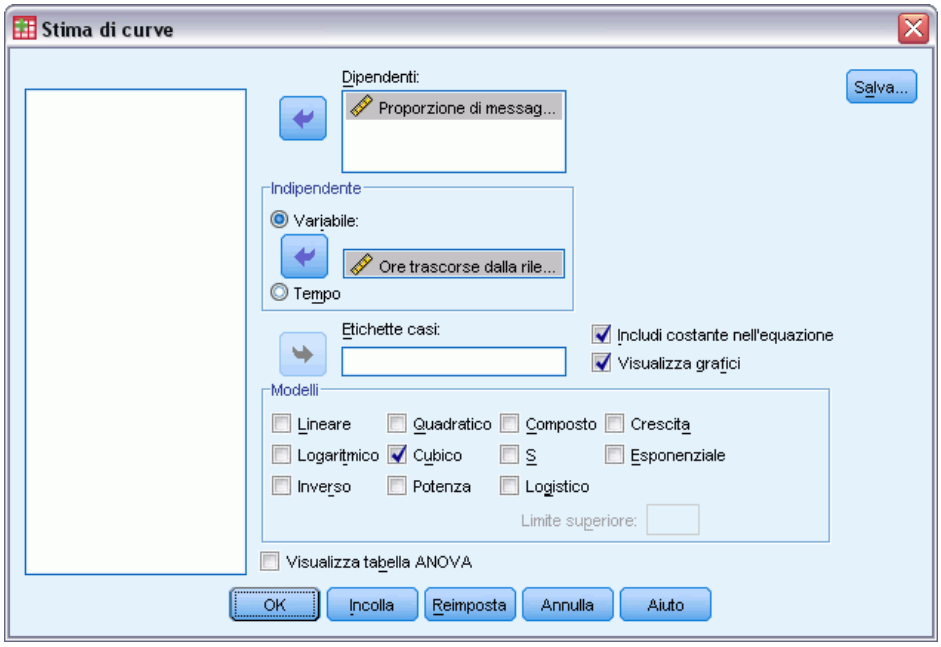
Dati. Le variabili dipendenti ed indipendenti devono essere quantitative. Se si seleziona Tempo come variabile indipenente dal file di dati attivo anziché selezionare una variabile, la procedura Stima di curve genera una variabile tempo se il periodo di tempo tra i casi è uniforme. Se si seleziona Tempo, la variabile dipendente deve essere una serie storica. L'analisi di serie storiche richiede nel file di dati una struttura in cui ciascun caso (riga) rappresenta una serie di osservazioni eseguite a orari diversi e il periodo di tempo fra i casi è uniforme.

Assunzioni. Valutare i dati graficamente per determinare il tipo di relazione esistente tra variabili dipendenti e indipendenti (lineare, esponenziale, ecc.). I residui di un buon modello devono essere normali e distribuiti casualmente. Se si utilizza un modello lineare, devono essere soddisfatte le seguenti ipotesi. Per ciascun valore della variabile indipendente, la distribuzione della variabile dipendente deve essere normale. La varianza della distribuzione della variabile dipendente deve essere costante per tutti i valori della variabile indipendente. La relazione tra la variabile dipendente e la variabile indipendente deve essere lineare e tutte le osservazioni devono essere indipendenti.

Per ottenere una stima di curve

- Dai menu, scegliere:
  Analizza > Regressione > Stima di curve...

Figura 18-1
Finestra di dialogo Stima di curve



- Selezionare una o più variabili dipendenti. Per ciascuna variabile dipendente verrà creato un modello distinto.
- Selezionare una variabile indipendente (una variabile nel file dati attivo oppure Tempo).
- Oppure:
  - Selezionare una variabile per etichettare casi nei grafici a dispersione. Per ciascun punto del grafico a dispersione, è possibile utilizzare lo strumento di selezione dei punti per visualizzare il valore della variabile Etichetta di caso.
  - Fare clic su Salva per salvare i valori attesi, i residui e gli intervalli di stima come nuove variabili.

Sono inoltre disponibili le seguenti opzioni:

- **Includi termine costante nell'equazione.** Consente di valutare un termine costante nell'equazione di regressione. La costante viene inclusa per impostazione predefinita.
- **Visualizza grafici.** Consente di tracciare i valori della variabile dipendente e ciascun modello selezionato in base alla variabile indipendente. Viene prodotto un grafico per ogni variabile dipendente.
- **Visualizza tabella ANOVA.** Consente di visualizzare una tabella di analisi della varianza per ciascun modello selezionato.

Stima di curve: Modelli

È possibile scegliere uno o più modelli di regressione per la stima di curve. Per determinare il modello da utilizzare, tracciare i dati in un grafico. Se le variabili appaiono legate da una relazione lineare, utilizzare un modello di regressione lineare semplice. Se la relazione tra le variabili non è lineare, provare a trasformare i dati. Se la trasformazione non risulta utile, può essere necessario utilizzare un modello più complesso. Visualizzare i dati in un grafico a dispersione; se il grafico è simile a una funzione matematica nota, adattare i dati a quel tipo di modello. Se, ad esempio, i dati sono simili a una funzione esponenziale, utilizzare il modello esponenziale.

Lineare. Modello la cui equazione è Y=b0+(b1*t). I valori della serie vengono rappresentati come una funzione lineare del tempo.

Logaritmico. Modello la cui equazione è Y = b0 + (b1 * ln(t)).

Inverso. Modello la cui equazione è Y = b0 + (b1 / t).

Quadratico. Modello la cui equazione è: Y=b0+(b1*t)+(b2*t**2). Il modello quadratico può essere usato per modellare una serie che "decolla" o una serie che si smorza rapidamente.

Cubico. Modello definito dall'equazione: Y = b0 + (b1 * t) + (b2 * t**2) + (b3 * t**3).

Potenza. Modello la cui equazione è Y = b0 * (t**b1) oppure ln(Y) = ln(b0) + (b1 * ln(t)).

Composto. Modello la cui equazione è Y = b0*(b1**t) oppure ln(Y) = ln(b0)+(ln(b1)*t).

Curva S. Modello la cui equazione è Y = e**(b0 + (b1/t)) oppure ln(Y) = b0 + (b1/t).

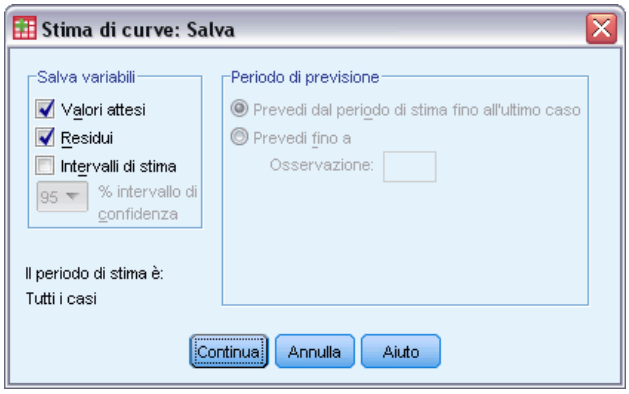
Logistica. Modello la cui equazione è Y=1/(1/u+(b0*(b1**t))) oppure ln(1/Y-1/u)=ln(b0+(ln(b1)*t)) dove u è il limite superiore. Per utilizzare l'equazione di regressione, specificare il valore limite superiore dopo aver selezionato Logistico. Tale valore deve essere un intero positivo maggiore del valore più alto assunto dalla variabile dipendente.

Crescita. Modello la cui equazione è Y = e**(b0 + (b1 * t)) o ln(Y) = b0 + (b1 * t).

Esponenziale. Modello la cui equazione è Y = b0 * (e**(b1 * t)) o ln(Y) = ln(b0) + (b1 * t).

Stima di curve: Salva

Figura 18-2
Finestra di dialogo Stima di curve: Salva



Salva Variabili. Per ciascun modello selezionato è possibile salvare i valori attesi, i residui (valori osservati della variabile dipendente meno il valore atteso del modello) e gli intervalli di stima (limite superiore e inferiore). I nomi delle nuove variabili e le etichette descrittive vengono visualizzati in una tabella nell'output.

Periodo di previsione. Se come variabile indipendente si seleziona Tempo anzichè una variabile nel file dati attivo, è possibile specificare un periodo di previsione al termine delle serie storiche. È possibile scegliere una delle seguenti opzioni:

- **Prevedi dal periodo di stima fino all'ultimo caso.** Consente di prevedere i valori per tutti i casi del file in base ai casi inclusi nel periodo di stima. Il periodo di stima, visualizzato in fondo alla finestra di dialogo, viene definito nella sottofinestra di dialogo dell'opzione Seleziona casi del menu Dati. Se non è stato definito alcun periodo di stima, i valori verranno previsti in base a tutti i casi.
- **Prevedi fino a.** Consente di prevedere i valori fino alla data, all'ora o al numero di osservazione specificato in base ai casi inclusi nel periodo di stima. Questa funzione può essere usata per prevedere valori futuri nelle serie storiche. Le variabili data definite specificano quali caselle di testo è possibile usare per specificare la fine di un periodo di previsione. Se non vengono definite variabili di data, è possibile specificare l'ultimo numero di osservazione (caso).

Per creare variabili di data, utilizzare l'opzione Definisci date, disponibile nel menu Dati.

Capitolo
19

Regressione minimi quadrati parziali

La procedura Regressione parziale minimi quadrati consente di stimare i modelli di regressione parziale dei minimi quadrati (PLS, nota anche come "proiezione della struttura latente"). PLS è una tecnica predittiva che rappresenta un'alternativa alla regressione dei minimi quadrati ordinari (OLS), alla correlazione canonica o ai modelli di equazioni strutturali e si rivela particolarmente utile quando le variabili predittore sono strettamente correlate o quando il numero dei predittori supera il numero dei casi.

PLS combina le funzioni dell'analisi dei componenti principali e della regressione multipla. Consente innanzitutto di estrarre un insieme di fattori latenti che forniscono la maggiore quantità di informazioni possibile sulla covarianza tra le variabili dipendenti e indipendenti. Quindi, un passo di regressione consente di prevedere i valori delle variabili dipendenti mediante la decomposizione delle variabili indipendenti.

Disponibilità. PLS è un comando di estensione che per la sua esecuzione richiede che nell'apposito sistema sia installato IBM® SPSS® Statistics - Integration Plug-In for Python. È necessario installare separatamente il modulo di estensione PLS, che può essere scaricato dall'indirizzo Web *http://www.ibm.com/developerworks/spssdevcentral*.

Tabelle. La proporzione della varianza spiegata (per fattore latente), i pesi fattoriali latenti, i carichi fattoriali latenti, l'importanza della variabile indipendente nella proiezione (VIP) e le stime dei parametri di regressione (per variabile dipendente) vengono tutti generati per impostazione predefinita.

Grafici. L'importanza della variabile nella proiezione (VIP), i punteggi fattoriali, i pesi fattoriali per i primi tre fattori latenti e la distanza dal modello vengono tutti generati dalla scheda Opzioni.

Livello di misurazione. Le variabili dipendenti e indipendenti (predittore) possono essere di scala, nominali o ordinali. La procedura presume che il livello di misurazione appropriato sia stato assegnato a tutte le variabili, sebbene sia possibile modificare temporaneamente il livello di misurazione di una variabile facendo clic con il pulsante destro del mouse sulla variabile nell'elenco delle variabili sorgente e scegliendo un livello di misurazione dal menu di scelta rapida. Le variabili categoriali (nominali o ordinali) vengono trattate in maniera equivalente dalla procedura.

Codifiche variabili categoriali. La procedura ricodifica temporaneamente le variabili dipendenti categoriali utilizzando le codifiche one-of-c*c* per la durata della procedura. Se esistono categorie *c* di una variabile, la variabile viene archiviata come vettori *c*, con la prima categoria indicata (1,0,...,0), la categoria successiva (0,1,0,...,0), ..., e la categoria finale (0,0,...,0,1). Le variabili dipendenti categoriali vengono rappresentate utilizzando una codifica fittizia; ovvero, semplicemente omettendo l'indicatore corrispondente alla categoria di riferimento.

Ponderazione. I valori dei pesi, prima di essere utilizzati vengono arrotondati ai numeri interi più vicini. I casi con pesi mancanti o con pesi inferiori a 0,5 non vengono utilizzati nelle analisi.

Valori mancanti. I valori mancanti di sistema e definiti dall'utente vengono considerati come non validi.

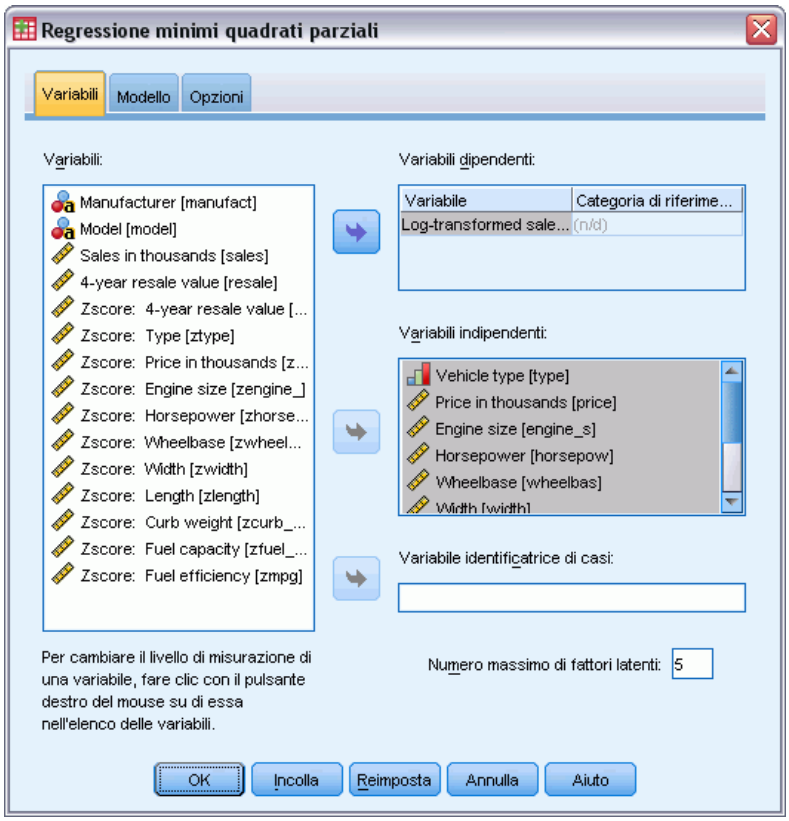
Modifica della scala. Tutte le variabili di modello sono centrate e standardizzate, comprese le variabili indicatore che rappresentano le variabili categoriali.

Per ottenere la regressione parziale dei minimi quadrati

Dai menu, scegliere:

Analizza > Regression > Minimi quadrati parziali...

Figura 19-1
Scheda Variabili della finestra Regressione parziale minimi quadrati



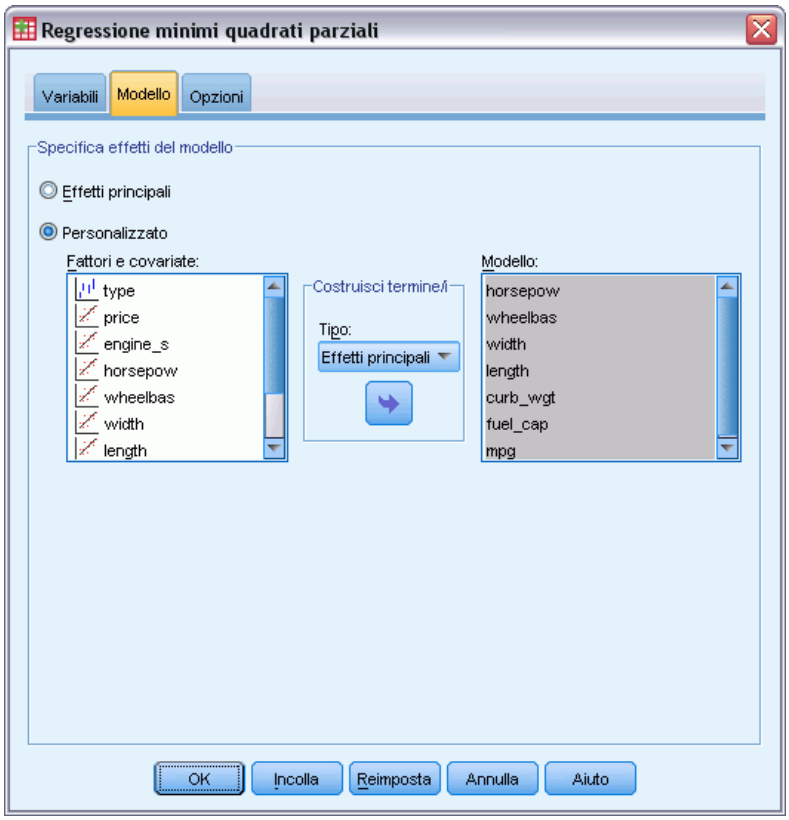
- Selezionare almeno una variabile dipendente.
- Selezionare almeno una variabile indipendente.

Se lo si desidera, è possibile:

- Specificare una categoria di riferimento per le variabili dipendenti categoriali (nominali o ordinali).
- Specificare una variabile da utilizzare come identificativo univoco per l'output per i casi e gli insiemi di dati salvati.
- Specificare un limite superiore per il numero dei fattori latenti da estrarre.

Modello

Figura 19-2
Scheda Modello della finestra Regressione parziale minimi quadrati



Specifica modello effetti. Un modello di effetti principali include tutti gli effetti principali di covariate e fattori. Selezionare Personalizzato per specificare le interazioni. È necessario indicare tutti i termini da includere nel modello.

Fattori e covariate. I fattori e le covariate sono elencati.

Modello. Il modello varia in base alla natura dei dati in uso. Dopo aver selezionato Personalizzato, è possibile selezionare gli effetti principali e le interazioni desiderate per l'analisi da eseguire.

Costruisci termini

Per i fattori e le covariate selezionati:

Interazione. Consente di creare il termine di interazione di livello maggiore rispetto a tutte le variabili selezionate. È l'impostazione di default.

Effetti principali. Consente di creare un termine di effetti principali per ciascuna variabile selezionata.

Tutti 2-vie. Consente di creare tutte le possibili interazioni a due vie delle variabili selezionate.

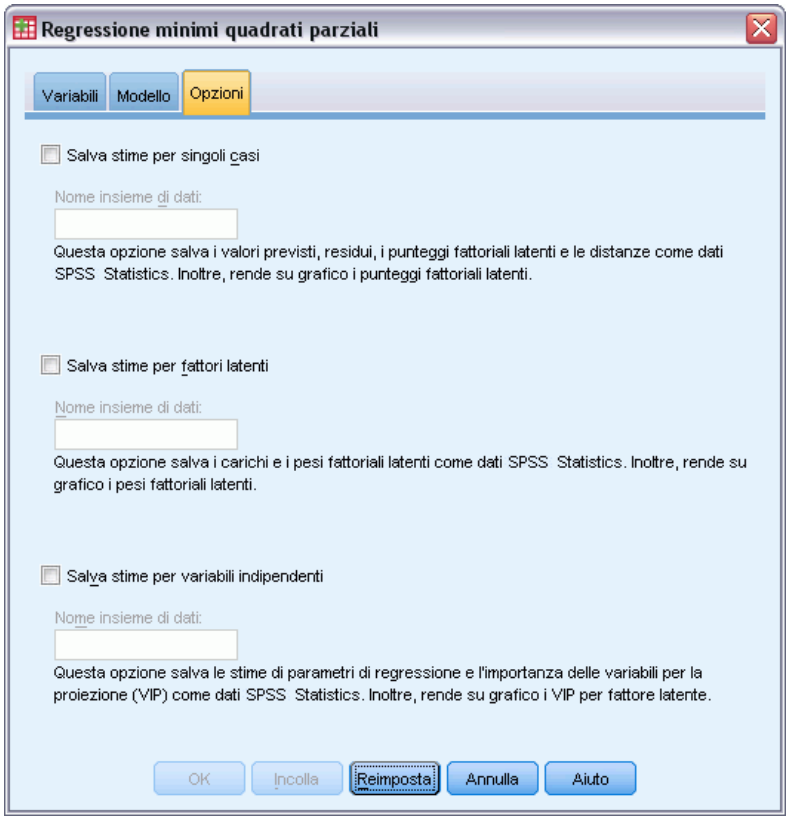
Tutti 3-vie. Consente di creare tutte le possibili interazioni a tre vie delle variabili selezionate.

Tutti 4-vie. Consente di creare tutte le possibili interazioni a quattro vie delle variabili selezionate.

Tutti 5-vie. Consente di creare tutte le possibili interazioni a cinque vie delle variabili selezionate.

Opzioni

Figura 19-3
Scheda Opzioni della finestra Regressione parziale minimi quadrati



La scheda Opzioni consente all'utente di salvare e rappresentare graficamente le stime dei modelli per singoli casi, fattori latenti e predittori.

Per ogni tipo di dati, specificare il nome di un insieme di dati. I nomi degli insiemi di dati devono essere univoci. Se si specifica il nome di un insieme di dati esistente, i suoi contenuti vengono sostituiti; altrimenti, viene creato un nuovo insieme di dati.

- **Salva stime per singoli casi.**Consente di salvare le stime dei modelli caso per caso, ovvero: i valori attesi, i residui, la distanza del modello dei fattori latenti e i punteggi fattoriali latenti. Inoltre, rende su grafico i punteggi fattoriali latenti.
- **Salva stime per fattori latenti.** Consente di salvare i carichi e i pesi fattoriali latenti. Inoltre, rende su grafico i pesi fattoriali latenti.
- **Salva stime per variabili indipendenti.** Consente di salvare le stime dei parametri di regressione e l'importanza delle variabili per la proiezione (VIP). Inoltre, rende su grafico i VIP per fattore latente.

Capitolo
20

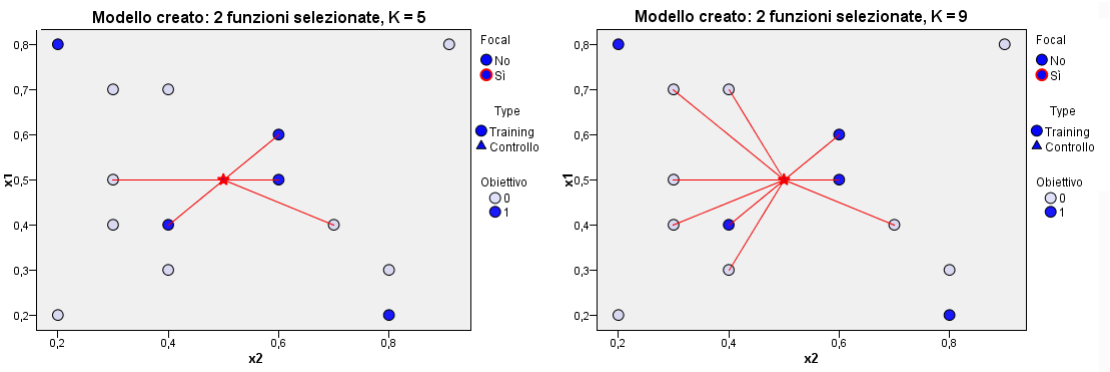
Analisi del vicino più vicino

L'analisi del vicino più vicino è un metodo per la classificazione dei casi basato sulla similarità ad altri casi. Nell'apprendimento automatico, questo metodo è stato sviluppato per riconoscere modelli di dati senza richiedere una corrispondenza esatta con eventuali modelli o casi archiviati. I casi simili sono vicini gli uni agli altri, mentre i casi dissimili sono distanti gli uni dagli altri. Pertanto, la distanza tra due casi rappresenta una misura della loro dissimilarità.

I casi attigui vengono definiti "vicini". Quando viene presentato un nuovo caso (controllo), viene calcolata la sua distanza da ogni caso del modello. Le classificazioni dei casi più simili, ovvero i vicini più vicini, vengono registrate e il nuovo caso viene inserito nella categoria contenente il numero più alto di vicini più vicini.

È possibile specificare il numero di vicini più vicini da esaminare; tale valore viene definito *k*. Le immagini mostrano in che modo viene classificato un nuovo caso utilizzando due valori diversi di *k*. Quando *k* = 5, il nuovo caso viene inserito nella categoria *1*, perché la maggioranza dei vicini più vicini appartiene alla categoria *1*. Tuttavia, quando *k* = 9, il nuovo caso viene inserito nella categoria *0*, perché la maggioranza dei vicini più vicini appartiene alla categoria *0*.

Figura 20-1
Effetti della modifica di k sulla classificazione



L'analisi dei vicini più vicini può essere utilizzata anche per calcolare i valori per un obiettivo continuo. In questa situazione, il valore di destinazione medio o mediano dei vicini più vicini viene utilizzato per ottenere il valore previsto per il nuovo caso.

Obiettivo e funzioni. L'obiettivo e le funzioni possono essere:

- **Nominale.** Una variabile può essere considerata nominale quando i relativi valori rappresentano categorie prive di ordinamento intrinseco, per esempio l'ufficio di una società, Tra gli esempi di variabili nominali troviamo la regione, il codice postale e la religione.

- **Ordinale.** Una variabile può essere considerata ordinale quando i relativi valori rappresentano categorie con qualche ordinamento intrinseco, per esempio i gradi di soddisfazione per un servizio, da molto insoddisfatto a molto soddisfatto, i punteggi di atteggiamento corrispondenti a gradi di soddisfazione o fiducia e i punteggi di preferenza.
- **Scala.** Una variabile può essere considerata di scala (continua) quando i relativi valori rappresentano categorie ordinate con una metrica significativa, tale che i confronti fra le distanze dei relativi valori siano appropriati. Esempi di variabili di scala sono l'età espressa in anni o il reddito espresso in migliaia di Euro.

  Le variabili nominali e ordinali vengono trattate in modo analogo dall'analisi del vicino più vicino. La procedura presume che il livello di misurazione appropriato sia stato assegnato a ciascuna variabile. Tuttavia, è possibile modificare temporaneamente il livello di misurazione di una variabile facendo clic con il pulsante destro del mouse sulla variabile nell'elenco delle variabili sorgente e scegliendo un livello di misurazione dal menu di scelta rapida.

L'icona accanto a ciascuna variabile nell'elenco delle variabili identifica il livello di misurazione e il tipo di dati.

	Numerico	Stringa	Data	Ora
Scala (continuo)		n/d		
Ordinale				
Nominale				

Codifiche variabili categoriali. La procedura ricodifica temporaneamente le variabili dipendenti e indipendenti categoriali utilizzando le codifiche one-of-*c* per la durata della procedura. Se esistono categorie *c* di una variabile, la variabile viene archiviata come vettori c, con la prima categoria indicata (1,0,...,0), la categoria successiva (0,1,0,...,0), ..., e la categoria finale (0,0,...,0,1).

Questo schema di codifica aumenta la dimensionalità dello spazio delle funzioni. In particolare, il numero totale di dimensioni è pari al numero di predittori di scala più il numero di categorie in tutti i predittori categoriali. Ne consegue che questo schema di codifica può generare un training più lento. Se il training del vicino più vicino sta procedendo molto lentamente, è possibile cercare di ridurre il numero di categorie nei predittori categoriali mediante la combinazione di categorie simili o casi di rilascio con categorie estremamente rare prima dell'esecuzione della procedura.

Tutta la codifica one-of-*c* è basata sui dati di training, anche se viene definito un campione di verifica o di controllo (vedere Partizioni). Pertanto, se il campione di controllo contiene casi con categorie di predittori assenti nei dati di training, non è possibile calcolarne il punteggio. Se il campione di controllo contiene casi con categorie di variabili dipendenti assenti nei dati di training, è invece possibile calcolarne il punteggio.

Modifica della scala. Le funzioni di scala vengono normalizzate per impostazione predefinita. Tale modifica viene eseguita interamente sulla base dei dati di training, anche se viene definito un campione di controllo (vedere Partizioni a pag. 134). Se si specifica una variabile per definire le partizioni, è importante che tali funzioni abbiano distribuzioni simili nei campioni di training

e controllo. Utilizzare, ad esempio, la procedura Esplora per esaminare le distribuzioni nelle partizioni.

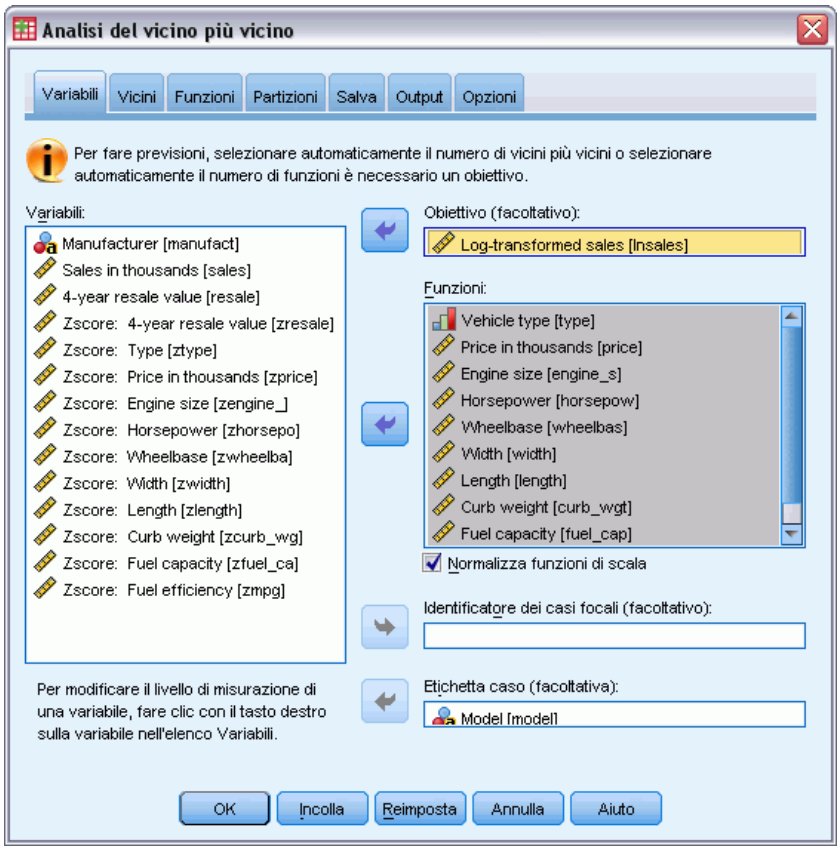
Ponderazione. La ponderazione viene ignorata da questa procedura.

Replica dei risultati. La procedura utilizza la generazione di numeri casuali durante l'assegnazione causale delle partizioni e dei sottocampioni con convalida incrociata. Se si desidera replicare esattamente i risultati ottenuti, oltre a utilizzare le stesse impostazioni per la procedura, è necessario impostare il valore di seme per il generatore di Mersenne Twister (vedere Partizioni a pag. 134) o utilizzare le variabili per definire le partizioni e i sottocampioni con convalida incrociata.

Per ottenere un'analisi del vicino più vicino

Dai menu, scegliere:
Analizza > Classifica > Vicino più vicino...

Figura 20-2
Finestra di dialogo Analisi del vicino più vicino, scheda Variabili



- Specificare una o più funzioni, che possono essere considerate variabili o predittori indipendenti in presenza di un obiettivo.

Obiettivo (facoltativo). Se non viene specificato alcun obiettivo (variabile o risposta dipendente), vengono rilevati esclusivamente i *k* vicini più vicini. Non vengono eseguite classificazioni né previsioni.

Normalizza funzioni di scala. Le funzioni normalizzate hanno lo stesso intervallo di valori, il che può migliorare le prestazioni dell'algoritmo di stima. Viene utilizzata la normalizzazione corretta, [2*(x−min)/(max−min)]−1. I valori normalizzati corretti sono compresi tra −1 e 1.

Identificatore dei casi focali (facoltativo). Consente di contrassegnare casi di particolare interesse. Si supponga, ad esempio, che un ricercatore desideri stabilire se i punteggi dei test di un determinato distretto scolastico (il caso focale) sono paragonabili a quelli di distretti scolastici analoghi. Ricorrerà all'analisi del vicino più vicino per individuare i distretti scolastici con le maggiori analogie in termini di uno specifico insieme di funzioni. Procederà quindi al confronto tra i punteggi dei test del distretto scolastico focale e quelli dei vicini più vicini.

I casi focali possono essere usati anche negli studi clinici per selezionare casi di controllo simili a casi clinici. I casi focali vengono visualizzati nella tabella dei *k* vicini più vicini e delle distanze, nel grafico dello spazio delle funzioni, in quello degli equivalenti e nella mappa dei quadranti. Le informazioni sui casi focali vengono salvate nei file specificati nella scheda Output.

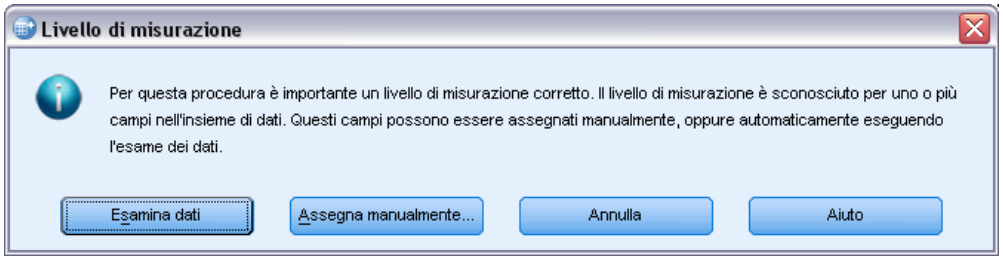
I casi con un valore positivo per la variabile specificata vengono trattati come casi focali. Non è consentito specificare una variabile senza valori positivi.

Etichetta caso (facoltativa). Ai casi vengono applicate etichette utilizzando questi valori nel grafico dello spazio di funzioni, in quello degli equivalenti e nella mappa dei quadranti.

Campi con livello di misurazione sconosciuto

L'avviso Livello di misurazione viene visualizzato quando il livello di misurazione di una o più variabili (campi) dell'insieme di dati è sconosciuto. Poiché influisce sul calcolo dei risultati di questa procedura, il livello di misurazione deve essere definito per tutte le variabili.

Figura 20-3
Avviso Livello di misurazione

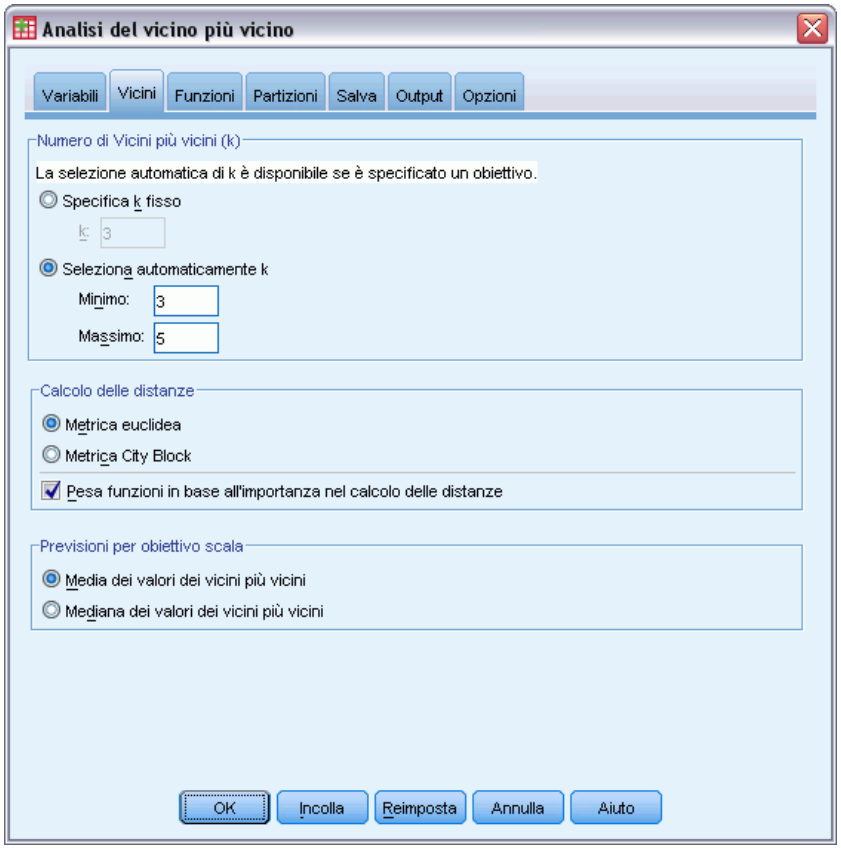


- **Esamina dati.** Legge i dati dell'insieme di dati attivo e assegna un livello di misurazione predefinito a tutti i campi con livello di misurazione sconosciuto. Con insiemi di dati di grandi dimensioni, questa operazione può richiedere del tempo.
- **Assegna manualmente.** Apre una finestra di dialogo che elenca tutti i campi con livello di misurazione sconosciuto, mediante la quale è possibile assegnare un livello di misurazione a questi campi. Il livello di misurazione si può assegnare anche nella Visualizzazione variabili dell'Editor dei dati.

Dal momento che il livello di misurazione è importante per questa procedura, è possibile accedere alla finestra di dialogo per la sua esecuzione solo quando per tutti i campi è stato definito un livello di misurazione.

Vicini

Figura 20-4
Finestra di dialogo Analisi del vicino più vicino, scheda Vicini



Numero di Vicini più vicini (k). Specificare il numero di vicini più vicini. L'utilizzo di un numero maggiore di vicini non garantisce necessariamente un modello più preciso.

Se nella tabella Variabili è specificato un obiettivo, in alternativa è possibile indicare un intervallo di valori e lasciare che sia la procedura a scegliere il "miglior" numero di vicini all'interno di tale intervallo. Il metodo utilizzato per stabilire il numero di vicini più vicini dipende dalla necessità o meno della selezione delle funzioni nella scheda Funzioni.

- Se la selezione delle funzioni è attiva, viene eseguita per ciascun valore di *k* nell'intervallo richiesto e viene selezionato il *k* (con il relativo insieme di funzioni) con il tasso di errore più basso (o l'errore più basso della somma dei quadrati se l'obiettivo è una scala).
- Se, invece, la selezione delle funzioni non è attiva, viene utilizzata la convalida incrociata con sottocampioni *V* per selezionare il "miglior" numero di vicini. Per informazioni sul controllo dell'assegnazione dei sottocampioni, vedere la scheda Partizione.

Calcolo delle distanze. Metrica utilizzata per specificare la metrica di distanza per la misurazione della similarità dei casi.

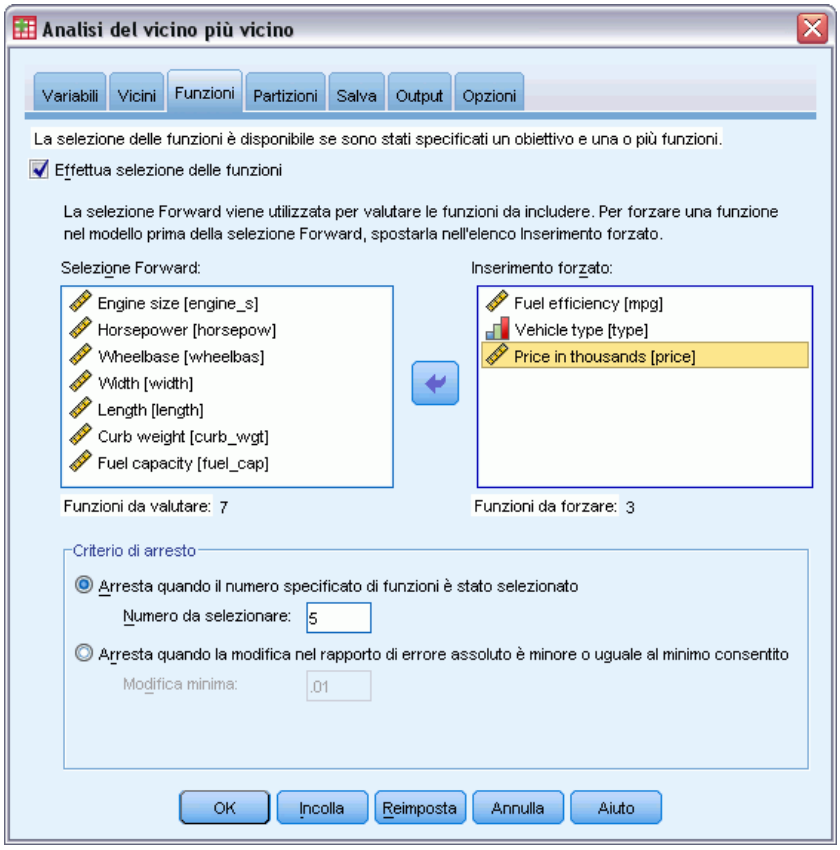
- **Metrica euclidea.** La distanza tra due casi, x e y, è pari alla radice quadrata della somma, in tutte le dimensioni, dei quadrati delle differenze tra i valori di tali casi.
- **Metrica city-block.** La distanza tra due casi è pari alla somma, in tutte le dimensioni, delle differenze assolute tra i valori di tali casi. È denominata anche "distanza di Manhattan".

Se si desidera, qualora nella scheda Variabili sia specificato un obiettivo, è possibile scegliere di ponderare le funzioni in base alla loro importanza normalizzata durante il calcolo delle distanze. L'importanza della funzione di un predittore si calcola dividendo il tasso di errore o l'errore della somma dei quadrati relativo al modello senza il predittore per il tasso di errore o l'errore della somma dei quadrati relativo al modello completo. L'importanza normalizzata si calcola riponderando i valori di importanza della funzione in modo che la somma sia pari a 1.

Previsioni per obiettivo scala. Se nella scheda Variabili è specificato un obiettivo scala, questo indica se il valore previsto è calcolato in base alla media o alla mediana dei vicini più vicini.

Funzioni

Figura 20-5
Finestra di dialogo Analisi del vicino più vicino, scheda Funzioni



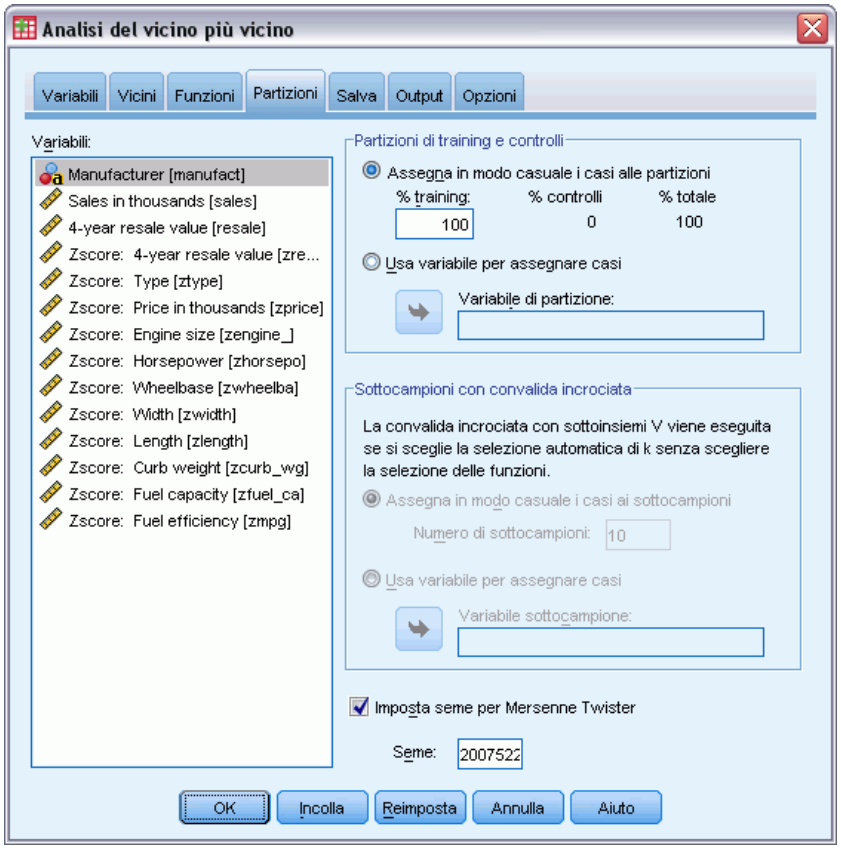
La scheda Funzioni consente di richiedere e specificare opzioni per la selezione delle funzioni quando nella scheda Variabili è specificato un obiettivo. Per impostazione predefinita, per la selezione delle funzioni vengono prese in considerazione tutte le funzioni, ma è possibile selezionare un sottoinsieme di funzioni da forzare nel modello.

Criterio di arresto. Di volta in volta, viene presa in considerazione, per essere inclusa nell'insieme dei modelli, la funzione il cui inserimento nel modello determina l'errore minore (calcolato come tasso di errore per gli obiettivi categoriali e come errore della somma dei quadrati per gli obiettivi scala). La selezione Forward prosegue fino al raggiungimento della condizione specificata.

- **Numero di funzioni specificato.** L'algoritmo inserisce un numero fisso di funzioni oltre a quelle forzate nel modello. Specificare un intero positivo. La riduzione dei valori del numero da selezionare dà origine a un modello più parsimonioso, con il rischio di perdere funzioni importanti. L'aumento dei valori del numero da selezionare consente di acquisire tutte le funzioni importanti, con il rischio però di aggiungere funzioni che finiscono per moltiplicare l'errore del modello.
- **Variazione minima nel rapporto di errore assoluto.** L'algoritmo si arresta quando la variazione del rapporto di errore assoluto indica che il modello non può essere migliorato ulteriormente aggiungendo altre funzioni. Specificare un numero positivo. La riduzione dei valori della variazione minima favorisce l'inclusione di un maggior numero di funzioni, con il rischio di inserire funzioni che non aggiungono valore al modello. L'aumento del valore della variazione minima, invece, tende a impedire l'inserimento di altre funzioni, con il rischio di perderne alcune importanti per il modello. Il valore "ottimale" della variazione minima dipende dai dati e dall'applicazione a disposizione. Per assistenza nella valutazione delle funzioni più importanti, vedere il registro degli errori relativi alla selezione delle funzioni nell'output. Per ulteriori informazioni, vedere l'argomento Registro degli errori relativi alla selezione delle funzioni a pag. 146.

Partizioni

Figura 20-6
Finestra di dialogo Analisi del vicino più vicino, scheda Partizioni



La scheda Partizioni consente di suddividere l'insieme di dati in sottoinsiemi di training e di controllo e, se possibile, di assegnare casi a sottocampioni con convalida incrociata.

Partizioni di training e di controllo. Questo gruppo specifica il metodo di partizionamento dell'insieme di dati attivo in campioni di training e di controllo. Il **campione di training** include i record di dati utilizzati per formare il modello di vicino più vicino; una percentuale di casi nell'insieme di dati deve essere assegnata al campione di training per ottenere un modello. Il **campione di controllo** è un insieme indipendente di record di dati utilizzato per valutare il modello finale; l'errore per il campione di controllo fornisce una stima "attendibile" della capacità predittiva del modello poiché i casi di controllo non sono stati utilizzati per generare il modello.

- **Assegna in modo casuale i casi alle partizioni.** Specificare la percentuale di casi da assegnare al campione di training. Il resto viene assegnato al campione di controllo.
- **Usa variabile per assegnare casi.** Specificare una variabile numerica che assegni ogni caso nell'insieme di dati attivo al campione di training o di controllo. I casi con valore positivo nella variabile vengono assegnati al campione di training, quelli con valore pari a 0 o negativo al campione di controllo. I casi con un valore di sistema mancante vengono esclusi dall'analisi. I valori mancanti definiti dall'utente per la variabile di partizione sono sempre considerati validi.

Sottocampioni con convalida incrociata. La convalida incrociata con sottocampioni *V* viene utilizzata per determinare il "miglior" numero di vicini. Per motivi legati alle prestazioni, la convalida incrociata non è disponibile se si utilizza la selezione delle funzioni.

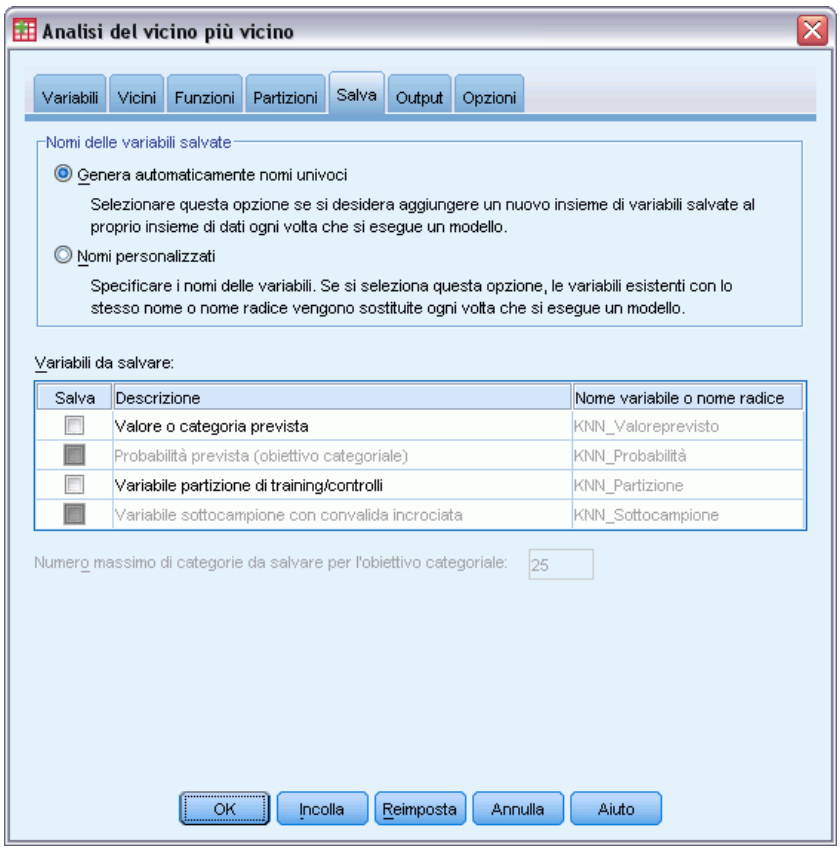
La convalida incrociata suddivide il campione in una serie di sottocampioni. I modelli di vicino più vicino vengono quindi generati escludendo di volta in volta i dati da ciascun sottocampione. Il primo modello si basa su tutti i casi eccetto quelli contenuti nel primo sottocampione, il secondo modello si basa su tutti i casi eccetto quelli contenuti nel secondo sottocampione e così via. Il rischio di errore per ciascun modello viene stimato applicando il modello al sottocampione escluso al momento della generazione del modello stesso. Il "miglior" numero di vicini più vicini è quello che genera l'errore più basso in tutti i sottocampioni.

- **Assegna in modo casuale i casi ai sottocampioni.** Specificare il numero di sottocampioni da utilizzare per la convalida incrociata. I casi vengono assegnati in modo casuale ai sottocampioni, numerati da 1 a *V*, il numero dei sottocampioni.
- **Usa variabile per assegnare casi.** Specificare una variabile numerica che assegni ogni caso nell'insieme di dati attivo a un sottocampione. La variabile deve essere un valore numerico compreso tra 1 e *V*. Se all'interno di tale intervallo mancano valori, e in corrispondenza delle distinzioni in caso di file distinti, si verificherà un errore.

Imposta seme per Mersenne Twister. Impostando un seme è possibile replicare le analisi. L'utilizzo di questo controllo è analogo all'impostazione di Mersenne Twister come generatore attivo specificando un punto di partenza fisso nella finestra di dialogo Generatori di numeri casuali, con un'importante differenza: impostando il seme in questa finestra di dialogo si conserva lo stato corrente del generatore di numeri casuali e lo si ripristina una volta terminata l'analisi.

Salva

Figura 20-7
Finestra di dialogo Analisi del vicino più vicino, scheda Salva



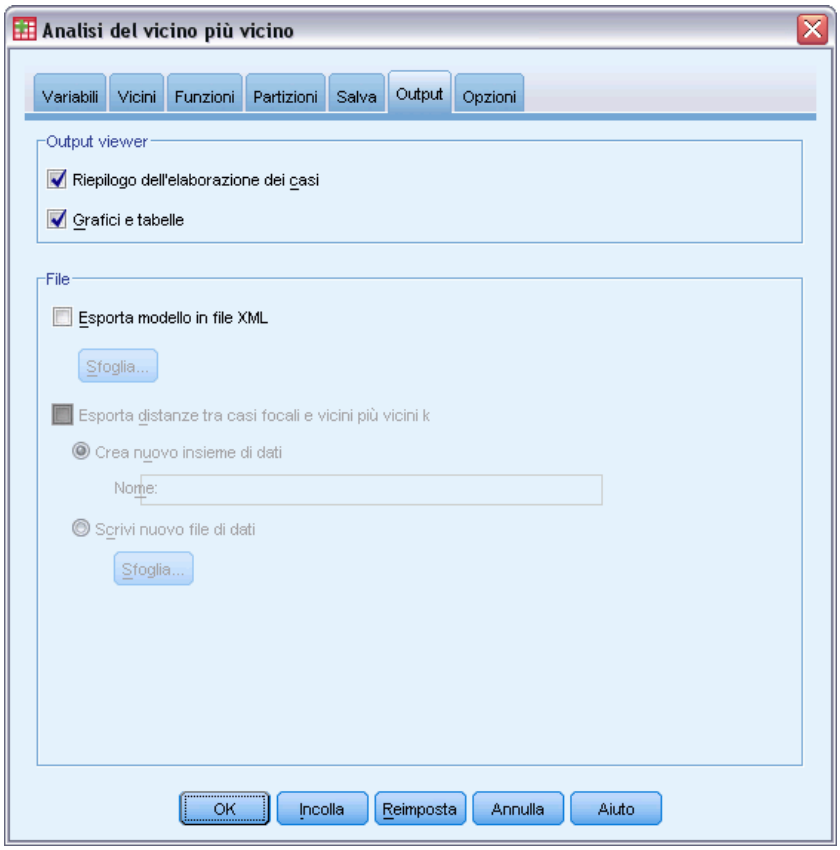
Nomi delle variabili salvate. La generazione automatica del nome assicura il mantenimento di tutto il lavoro. I nomi personalizzati consentono di eliminare/sostituire i risultati di precedenti esecuzioni senza dover prima eliminare le variabili salvate nell'Editor dei dati.

Variabili da salvare

- **Valore o categoria prevista.** Viene salvato il valore previsto per un obiettivo scala o la categoria prevista per un obiettivo categoriale.
- **Probabilità prevista.** Vengono salvate le probabilità previste per un obiettivo categoriale. Una variabile separata viene salvata per ognuna delle prime *n* categorie, dove *n* viene specificato nel comando Numero massimo di categorie da salvare per l'obiettivo categoriale.
- **Variabile partizione di training/controlli.** Se i casi vengono assegnati in modo casuale ai campioni training e di controllo nella scheda Partizioni, viene salvato il valore della partizione (di training o di controllo) a cui il caso è stato assegnato.
- **Variabile sottocampione con convalida incrociata.** Se nella scheda Partizioni ai sottocampioni con convalida incrociata vengono assegnati casi in modo casuale, viene salvato il valore del sottocampione a cui è stato assegnato il caso.

Output

Figura 20-8
Finestra di dialogo Analisi del vicino più vicino, scheda Output



Output viewer

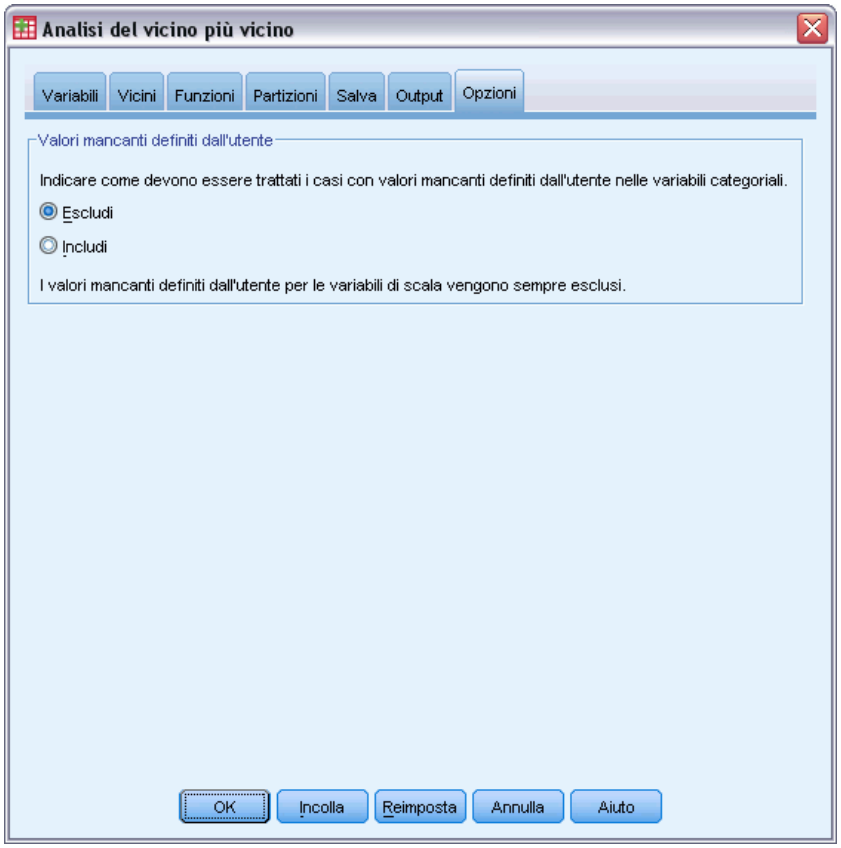
- **Riepilogo dell'elaborazione dei casi.** Visualizza la tabella di riepilogo di elaborazione dei casi, che riepiloga il numero di casi inclusi ed esclusi dall'analisi, in totale e per campioni di training e di controllo.
- **Grafici e tabelle.** Visualizza l'output relativo al modello, tra cui tabelle e grafici. Le tabelle nella vista del modello comprendono i vicini più vicini *k* e le distanze per i casi focali, la classificazione delle variabili di risposta categoriali e un riepilogo degli errori. L'output grafico nella vista del modello include un registro degli errori relativi alla selezione, il grafico dell'importanza delle funzioni, quello dello spazio di funzioni e degli equivalenti e la mappa dei quadranti. Per ulteriori informazioni, vedere l'argomento Vista del modello a pag. 139.

File

- **Esporta modello in XML.**È possibile utilizzare questo file di modello per applicare le informazioni del modello ad altri file di dati per il calcolo del punteggio. Questa opzione non è disponibile se sono stati definiti file suddivisi.
- **Esporta distanze tra casi focali e vicini più vicini k.** Per ogni caso focale viene creata una variabile distinta per ciascuno dei vicini più vicini *k* (dal campione di training) del caso focale stesso e delle corrispondenti distanze più vicine *k*.

Opzioni

Figura 20-9
Finestra di dialogo Analisi del vicino più vicino, scheda Output

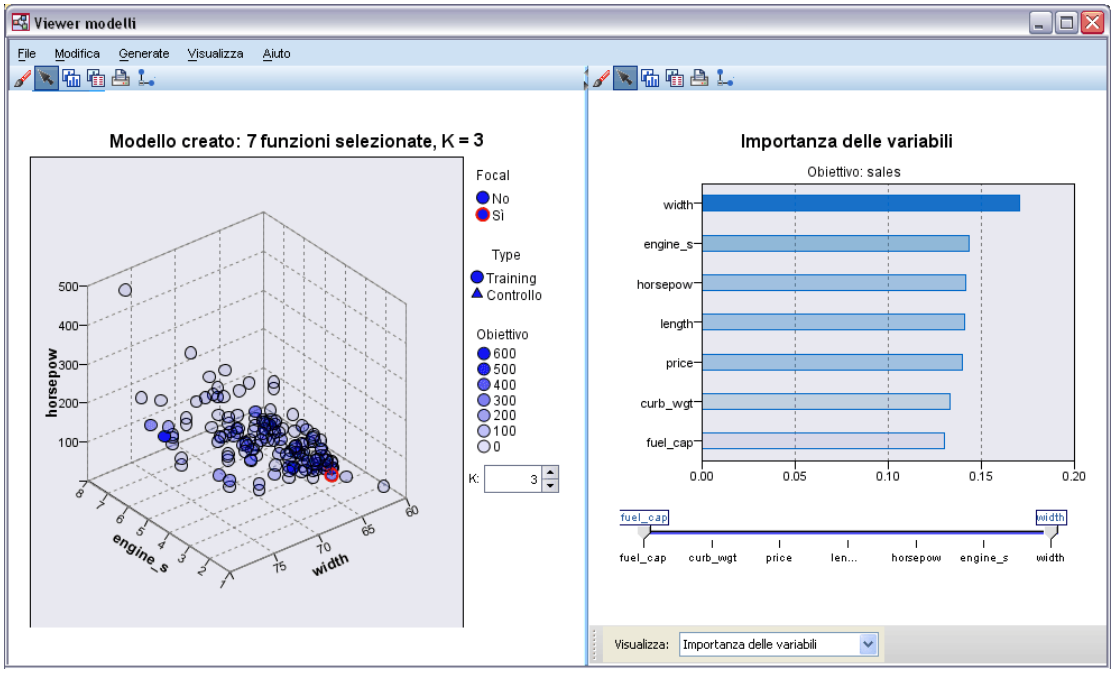


Valori mancanti definiti dall'utente. Le variabili categoriali devono contenere valori validi per un caso per essere incluse nell'analisi. Questi controlli consentono di decidere se i valori mancanti definiti dall'utente devono essere considerati validi nell'ambito delle variabili categoriali.

I valori mancanti di sistema e i valori mancanti relativi alle variabili di scala vengono sempre considerati non validi.

Vista del modello

Figura 20-10
Vista del modello nell'analisi del vicino più vicino



Selezionando Grafici e tabelle nella scheda Output, nel Viewer viene creato un oggetto Modello vicino più vicino. Attivando l'oggetto con un doppio clic, si accede a una vista interattiva del modello con una finestra a due riquadri:

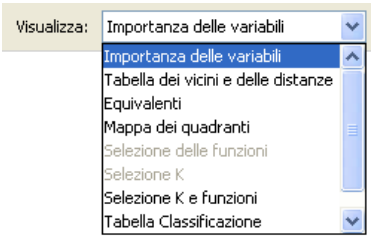
- Nel primo è presente una panoramica del modello denominata "vista principale".
- Nel secondo, invece, possono essere visualizzate due tipologie di vista:

  La vista ausiliaria mostra ulteriori informazioni sul modello, pur non concentrandosi su quest'ultimo.

  La vista collegata mostra invece i dettagli relativi a una funzione del modello quando l'utente esegue il drill-down di parte della vista principale.

Per impostazione predefinita, nel primo riquadro viene visualizzato lo spazio di funzioni e nel secondo il grafico dell'importanza delle variabili. Se quest'ultimo grafico non è disponibile (se, cioè, nella scheda Funzioni non è stato selezionato Pesa funzioni in base all'importanza), viene visualizzata la prima vista presente nell'elenco a discesa Vista.

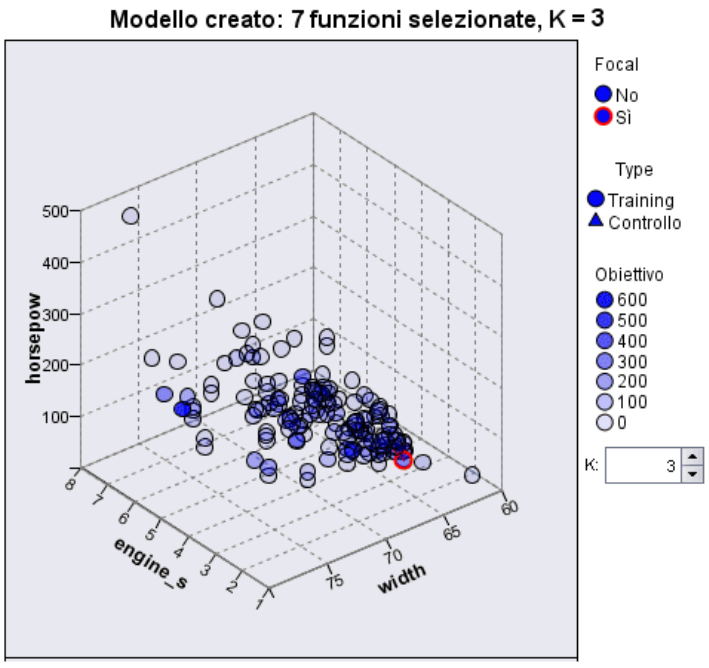
Figura 20-11
Finestra di dialogo Analisi del vicino più vicino, elenco a discesa Vista del modello



Quando per una vista non sono disponibili informazioni, la voce corrispondente nell'elenco a discesa Vista viene disattivata.

Spazio di funzioni

Figura 20-12
Spazio di funzioni



Il grafico dello spazio di funzioni è un grafico interattivo relativo allo spazio delle funzioni (o al sottospazio, se sono presenti più di tre funzioni). Ogni asse rappresenta una funzione nel modello e la posizione dei punti nel grafico indica i valori di tali funzioni per i casi nelle partizioni di training e di controllo.

Chiavi. Oltre a rappresentare i valori delle funzioni, i punti forniscono altre informazioni.

- La forma indica la partizione (Training o Controllo) di cui fa parte un punto.

- Il colore/l'ombreggiatura di un punto indica il valore dell'obiettivo del caso (i diversi valori di colore corrispondono alle categorie di un obiettivo categoriale, mentre le ombreggiature indicano l'intervallo di valori di un obiettivo continuo). Il valore indicato per la partizione di training è quello osservato, mentre per la partizione di controllo è indicato quello previsto. Se non viene specificato alcun obiettivo, questo simbolo non viene visualizzato.
- Contorni più marcati indicano che un caso è focale. I casi focali vengono visualizzati collegati ai relativi vicini più vicini *k*.

Comandi e interattività. Nel grafico è disponibile una serie di comandi per esplorare lo spazio di funzioni.

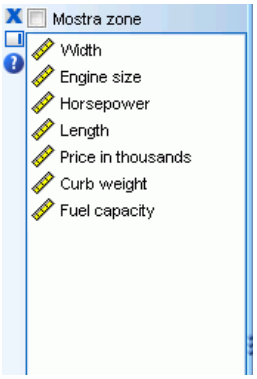
- È possibile scegliere il sottoinsieme di funzioni da visualizzare nel grafico e cambiare le funzioni da rappresentare nelle dimensioni.
- I casi focali non sono altro che punti selezionati nel grafico relativo allo spazio di funzioni. Se è stata specificata una variabile per casi focali, inizialmente verranno selezionati i punti che rappresentano i casi focali. Qualsiasi punto può comunque diventare temporaneamente un caso focale se viene selezionato. Vengono utilizzati i "soliti" comandi per la selezione di punti (facendo clic su un punto, quest'ultimo viene selezionato e vengono deselezionati tutti gli altri; facendo clic su un punto mentre si tiene premuto il tasto CTRL, tale punto viene aggiunto all'insieme dei punti selezionati). Le viste collegate, ad esempio il grafico degli equivalenti, vengono automaticamente aggiornate in base ai casi selezionati nello spazio di funzioni.
- È possibile modificare il numero di vicini più vicini (*k*) da visualizzare per i casi focali.
- Passando il mouse sopra un punto del grafico, viene visualizzata una descrizione con il valore dell'etichetta del caso (o il numero del caso se non sono state definite etichette), oltre ai valori osservati e previsti dell'obiettivo.
- Il pulsante di ripristino consente di reimpostare lo spazio di funzioni allo stato originario.

Aggiunta e rimozione di campi/variabili

Nello spazio di funzioni è possibile aggiungere nuovi campi/variabili o rimuovere quelli già visualizzati.

Tavolozza delle variabili

Figura 20-13
Tavolozza delle variabili



Per poter aggiungere e rimuovere le variabili è necessario prima visualizzare la tavolozza delle variabili. Per poterla visualizzare è necessario che il Viewer modelli sia in modalità Modifica e che sia selezionato un caso nello spazio di funzioni.

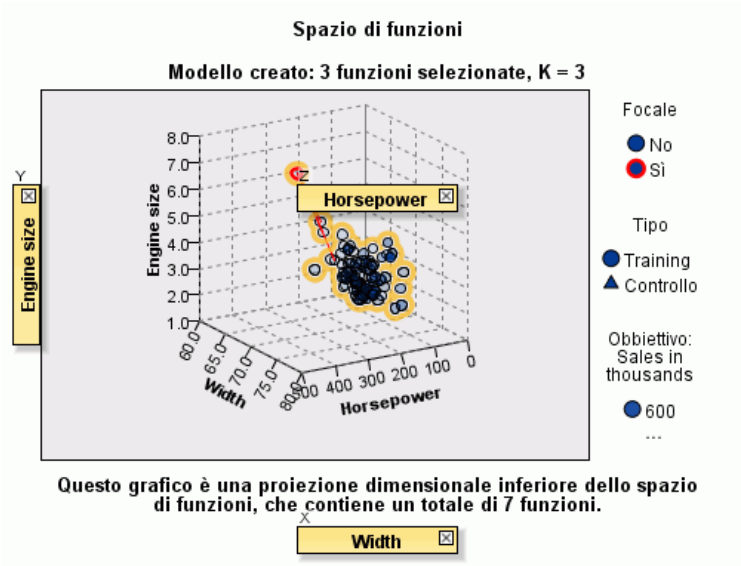
- Per attivare la modalità Modifica nel Viewer modelli, dai menu scegliere:
  Visualizza > Modalità Modifica
- Una volta in modalità Modifica, fare clic su un caso nello spazio di funzioni.
- Per visualizzare la tavolozza delle variabili, dai menu scegliere:
  Visualizza > Tavolozze > Variabili

  La tavolozza delle variabili elenca tutte le variabili presenti nello spazio di funzioni. L'icona accanto al nome della variabile indica il livello di misurazione della variabile.
- Per modificare temporaneamente il livello di misurazione di una variabile, fare clic con il pulsante destro del mouse nella tavolozza delle variabili e selezionare un'opzione.

Aree delle variabili

Le variabili vengono aggiunte a delle "aree" all'interno dello spazio di funzioni. Per visualizzare le aree, iniziare a trascinare una variabile dalla tavolozza delle variabili o selezionare Mostra zone.

Figura 20-14
Aree delle variabili



Lo spazio di funzioni dispone di aree per gli assi *x*, *y* e *z*.

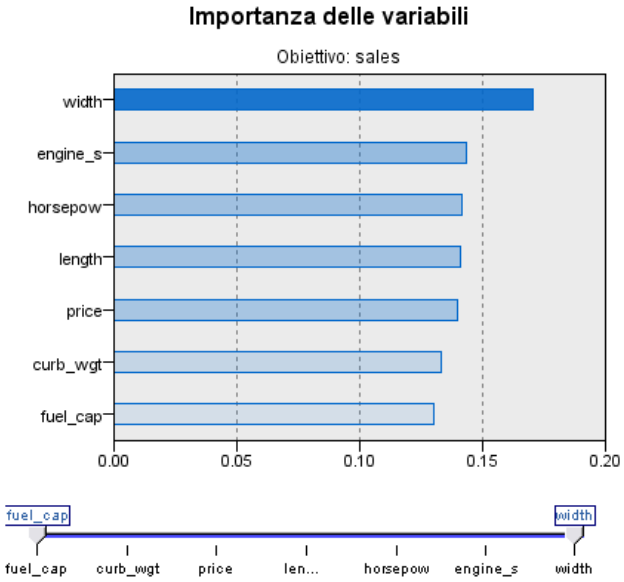
Spostamento delle variabili nelle aree

Di seguito sono riportate alcune regole e suggerimenti generali per spostare le variabili nelle aree:

- Per spostare una variabile in un'area, trascinare la variabile dalla tavolozza delle variabili all'interno dell'area. Se si seleziona Mostra zone è possibile anche fare clic con il pulsante destro del mouse su un'area e selezionare una variabile da aggiungere a quell'area.
- Se si trascina una variabile dalla tavolozza delle variabili a un'area già occupata da un'altra variabile, la vecchia variabile viene sostituita con la nuova.
- Se si trascina una variabile da un'area a un'area già occupata da un'altra variabile, le variabili si scambiano di posizione.
- Se si fa clic sulla X all'interno di un'area si rimuove la variabile presente in quell'area.
- Se la visualizzazione comprende più elementi grafici, ciascuno di essi può avere delle aree delle variabili associate. Selezionare prima l'elemento grafico desiderato.

Importanza della variabile

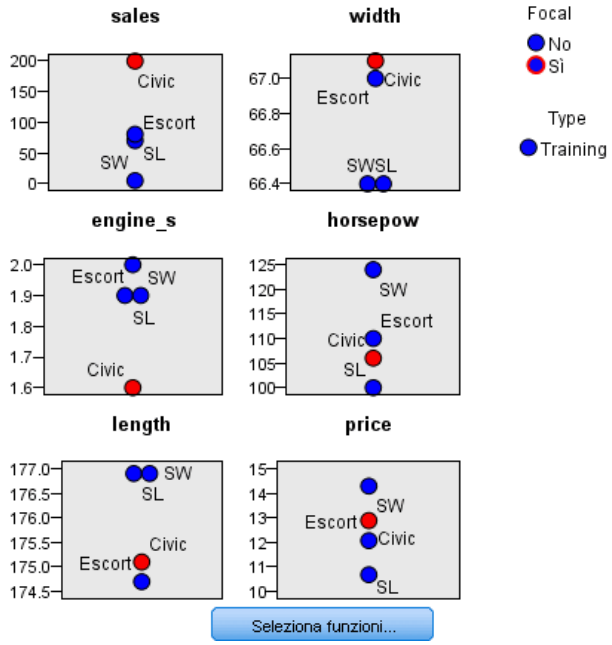
Figura 20-15
Importanza delle variabili



Di solito è opportuno concentrare la modellazione sulle variabili più rilevanti, lasciando perdere o ignorando le meno importanti. In questo senso può essere utile il grafico dell'importanza delle variabili, che indica l'importanza relativa di ciascuna variabile nella stima del modello. Dal momento che i valori sono relativi, la somma dei valori di tutte le variabili visualizzate è pari a 1,0. L'importanza delle variabili non ha nulla a che vedere con la precisione del modello. Riguarda unicamente l'importanza di ciascuna variabile per l'elaborazione di una previsione, non il grado di precisione di quest'ultima.

Equivalenti

Figura 20-16
Grafico degli equivalenti



In questo grafico vengono visualizzati i casi focali e i relativi vicini più vicini *k* per ciascuna funzione e per l'obiettivo. È disponibile se nello spazio di funzioni è selezionato un caso focale.

Collegamenti. Il grafico degli equivalenti è collegato allo spazio di funzioni in due modi.

- I casi selezionati (focali) nello spazio di funzioni vengono visualizzati nel grafico degli equivalenti, insieme ai relativi vicini più vicini *k*.
- Il valore di *k* selezionato nello spazio di funzioni viene utilizzato nel grafico degli equivalenti.

Distanze dei vicini più vicini

Figura 20-17
Distanze dei vicini più vicini

Caso focale	Vicini più vicini			Distanze più prossime		
	1	2	3	1	2	3
Civic	SL	Escort	SW	0.053	0.0599	0.0643

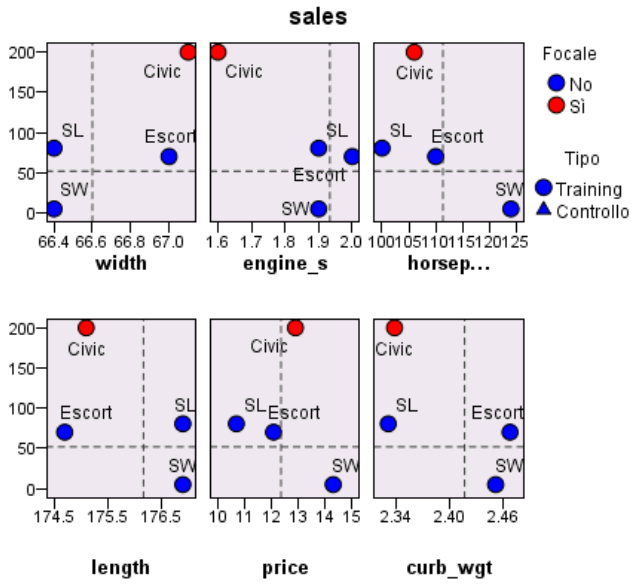
In questa tabella vengono visualizzati i vicini più vicini e le distanze più vicine *k* solo per i casi focali. È disponibile se nella scheda Variabili è specificato un identificatore dei casi focali e mostra soltanto i casi focali identificati da questa variabile.

Ogni riga della:

- Colonna Caso focale contiene il valore della variabile di etichetta relativa al caso focale. Se non sono definite etichette dei casi, la colonna contiene il numero di caso del caso focale.
- *i*ª colonna del gruppo Vicini più vicini contiene il valore della variabile di etichetta dei casi relativa al *i*° vicino più vicino del caso focale. Se non sono definite etichette dei casi, la colonna contiene il numero di caso del *i*° vicino più vicino del caso focale.
- *i*ª colonna del gruppo Distanze più vicine contiene la distanza del *i*° vicino più vicino dal caso focale.

Mappa dei quadranti

Figura 20-18
Mappa dei quadranti

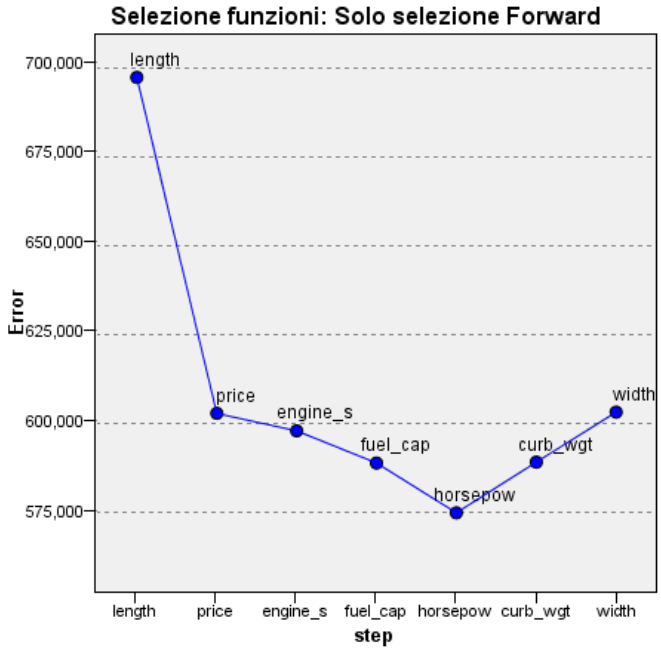


Il grafico mostra i casi focali e i relativi vicini più vicini *k* su un grafico a dispersione (o un grafico a punti a seconda del livello di misurazione dell'obiettivo) con l'obiettivo sull'asse *y* e una funzione di scala sull'asse *x*, il tutto suddiviso in riquadri in base alle funzioni. È disponibile se nello spazio di funzioni è presente un obiettivo ed è selezionato un caso focale.

- Per le variabili continue, nella partizione di training in corrispondenza delle medie delle variabili vengono tracciate linee di riferimento.

Registro degli errori relativi alla selezione delle funzioni

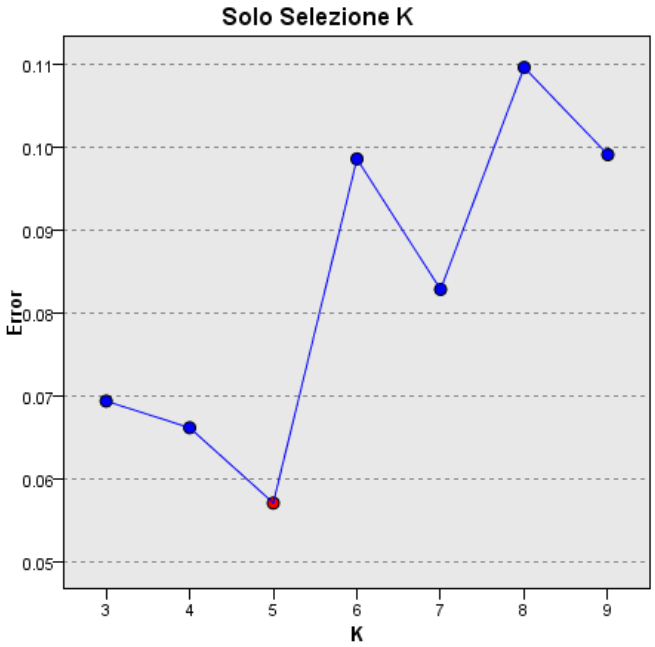
Figura 20-19
Selezione funzioni



I punti presenti nel grafico mostrano l'errore (in termini di tasso di errore o di errore della somma dei quadrati a seconda del livello di misurazione dell'obiettivo) sull'asse *y* del modello, con la funzione elencata sull'asse *x* (inoltre, a sinistra sull'asse *x* sono presenti tutte le funzioni). Il grafico è disponibile se è presente un obiettivo ed è attiva la selezione funzioni.

Registro degli errori relativi alla selezione di k

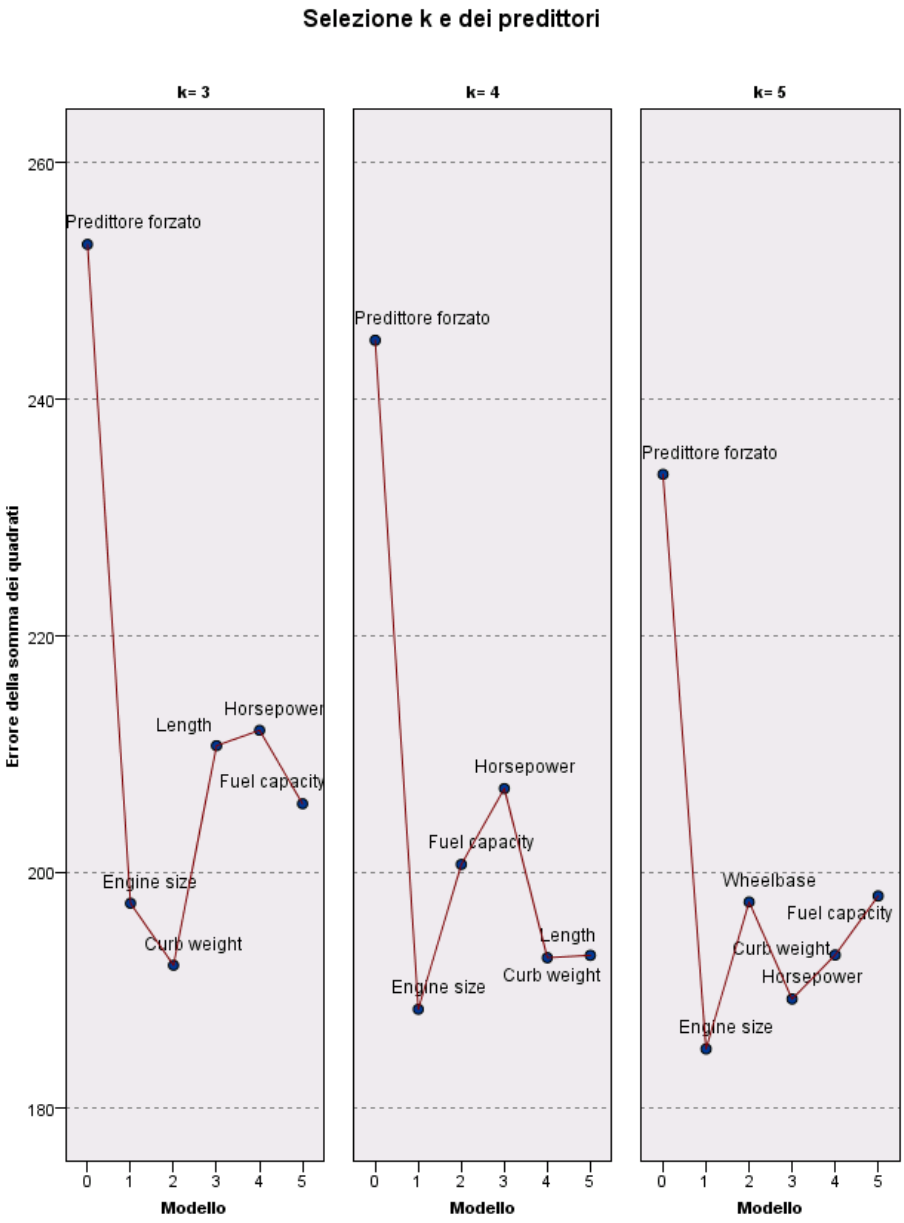
Figura 20-20
Selezione k



I punti presenti nel grafico mostrano l'errore (in termini di tasso di errore o di errore della somma dei quadrati a seconda del livello di misurazione dell'obiettivo) sull'asse *y* del modello, con il numero di vicini più vicini (*k*) sull'asse *x*. Il grafico è disponibile se è presente un obiettivo ed è attiva la selezione *k*.

Registro degli errori relativi alla selezione k e alla selezione delle funzioni

Figura 20-21
Selezione k e selezione delle funzioni



Si tratta di grafici per la selezione delle funzioni (vedere Registro degli errori relativi alla selezione delle funzioni a pag. 146), suddivisi in riquadri in base a *k*. Il grafico è disponibile se è presente un obiettivo e sono attive sia la selezione *k* sia la selezione delle funzioni.

Tabella di classificazione

Figura 20-22
Tabella di classificazione

Partizione		Previsto		
		0	1	Percentuale corretta
Training	0	111	1	99.11%
	1	7	33	82.50%
	Percentuale globale	77.64%	22.37%	94.74%

Nella tabella viene visualizzata la classificazione incrociata dei valori osservati dell'obiettivo rispetto a quelli previsti, suddivisi per partizione. È disponibile se è presente un obiettivo di tipo categoriale.

- La riga (Mancante) della partizione di controllo contiene casi di controllo con valori mancanti sull'obiettivo. Tali casi contribuiscono ai valori di percentuale complessiva, ma non a quelli di percentuale corretta, del campione di controllo.

Riepilogo degli errori

Figura 20-23
Riepilogo degli errori

Partizione	Sum-of-Squares Error
Training	622043

La tabella è disponibile in presenza di una variabile di destinazione. Mostra l'errore associato al modello: la somma dei quadrati per l'obiettivo continuo e il tasso di errore (percentuale complessiva di correttezza del 100%−) per un obiettivo categoriale.

Capitolo
21

Analisi discriminante

L'analisi discriminante crea un modello di previsione per il gruppo di appartenenza. Il modello è costituito da una funzione discriminante oppure, per più di due gruppi, da un insieme di funzioni discriminanti, in base alle combinazioni lineari delle variabili predittore che forniscono la migliore discriminazione tra i gruppi. Le funzioni vengono generate da un campione di casi di cui è noto il gruppo di appartenenza; è quindi possibile applicare le funzioni ai nuovi casi con misurazioni per le variabili predittore, ma di cui non è noto il gruppo di appartenenza.

Nota: La variabile di raggruppamento può includere più di due valori. I codici per la variabile di raggruppamento devono tuttavia essere interi, ed è necessario specificare i valori massimo e minimo corrispondenti. I casi con valore non compreso tra i due estremi specificati vengono esclusi dall'analisi.

Esempio. In media, il consumo calorico giornaliero degli abitanti delle zone temperate è maggiore di quello di chi vive ai tropici. Nelle zone temperate, inoltre, si riscontra una maggiore percentuale di persone che vivono in ambiente urbano. Un ricercatore desidera combinare queste informazioni in una funzione per determinare le modalità di discriminazione tra i due gruppi di paesi. Il ricercatore ritiene opportuno prendere in considerazione anche le dimensioni della popolazione e informazioni di carattere economico. L'analisi discriminante consente di valutare i coefficienti della funzione discriminante lineare, analoga alla parte destra di un'equazione di regressione lineare multipla. In altri termini, utilizzando i coefficienti *a*, *b*, *c* e *d* si ottiene la funzione:

```
D = a * clima + b * urbano + c * popolazione + d * prodotto interno lordo pro capite
```

Se queste variabili sono utili per la discriminazione tra le due zone climatiche, i valori di *D* per i paesi temperati saranno diversi da quelli relativi ai paesi tropicali. Se è necessario usare un metodo di selezione delle variabili per passi, nella funzione non si dovranno includere tutte e quattro le variabili.

Statistiche. Per ogni variabile: media, deviazione standard, ANOVA univariata. Per ogni analisi: *M* di Box, matrice di correlazione entro gruppi, matrice di covarianza entro gruppi, matrice di covarianza di gruppi separati e matrice di covarianza totale. Per ciascuna funzione discriminante canonica: autovalori, percentuale di varianza, correlazione canonica, lambda di Wilks, chi-quadrato. Per ogni passo: probabilità a priori, coefficienti di funzione di Fisher, coefficienti di funzione standardizzati, lambda di Wilks per ogni funzione canonica.

Dati. La variabile di raggruppamento deve includere un numero limitato di categorie distinte, codificate come interi. Le variabili indipendenti nominali devono essere ricodificate in forma di variabili fittizie o di contrasto.

Assunzioni. I casi devono essere indipendenti. Le variabili stimatore devono avere una distribuzione normale multivariata e le matrici di varianza-covarianza entro gruppi devono essere uguali in tutti i gruppi. Si assume che le appartenenze ai gruppi si escludano reciprocamente (ovvero che nessun caso appartenga a più gruppi) e che ciascun caso appartenga a un gruppo.

La procedura è più efficace se l'appartenenza ai gruppi è una variabile categoriale effettiva; se l'appartenenza ai gruppi si basa sui valori di una variabile continua (ad esempio, QI massimo e QI minimo), è opportuno utilizzare la regressione lineare per avvalersi delle informazioni più dettagliate disponibili nella variabile continua.

Per ottenere un'analisi discriminante

▶ Dai menu, scegliere:
Analizza > Classifica > Discriminante...

Figura 21-1
Finestra di dialogo Analisi discriminante

▶ Selezionare una variabile di raggruppamento con valori interi e fare clic su Definisci intervallo per specificare le categorie desiderate.

▶ Selezionare le variabili indipendenti o le variabili stimatore. (Se la variabile di raggruppamento non include valori interi, utilizzando il comando Ricodifica automatica del menu Trasforma è possibile crearne una che includa tali valori).

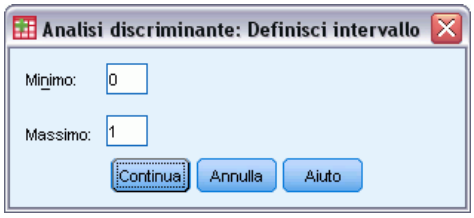
▶ Selezionare il metodo di inserimento delle variabili indipendenti.

- **Inserisci indipendenti insieme.** Inserisce contemporaneamente tutte le variabili indipendenti che soddisfano i criteri di tolleranza.
- **Usa metodo stepwise.** Usa un metodo per passi per controllare l'inserimento e la rimozione delle variabili.

▶ È inoltre possibile selezionare i casi utilizzando una variabile di selezione.

Analisi discriminante: Definisci intervallo

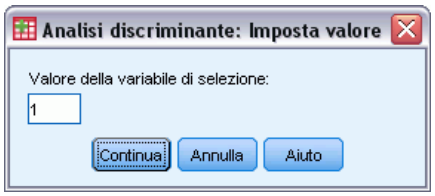
Figura 21-2
Finestra di dialogo Analisi discriminante: Definisci intervallo



Specificare il valore minimo e massimo della variabile di raggruppamento da utilizzare per l'analisi. I casi con valori che non rientrano in tale intervallo non vengono utilizzati nell'analisi discriminante, ma vengono classificati in uno dei gruppi esistenti in base ai risultati dell'analisi stessa. I valori massimo e minimo devono essere interi.

Analisi discriminante: Seleziona casi

Figura 21-3
Finestra di dialogo Analisi discriminante: Imposta valore



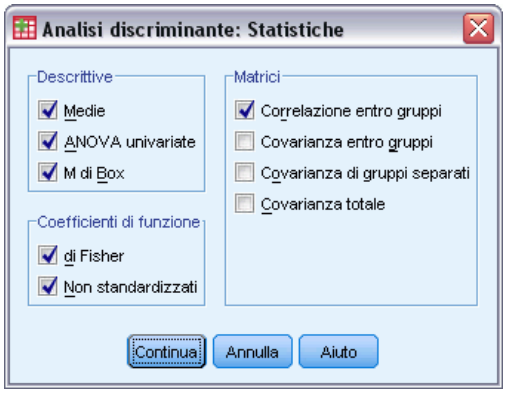
Per selezionare i casi da usare nell'analisi:

- Scegliere la variabile di selezione nella finestra di dialogo Analisi discriminante.
- Fare clic suValore per immettere la variabile di selezione come numero intero.

Le funzioni discriminanti verranno derivate solo in base ai casi che prevedono tale valore per la variabile di selezione. Le statistiche e i risultati della classificazione vengono generati sia per i casi selezionati che per i casi non selezionati. Tale processo consente di classificare i nuovi casi sulla base dei dati preesistenti nonché ripartire i dati in sottoinsiemi di prova con i quali eseguire la convalida del modello generato.

Analisi discriminante: Statistiche

Figura 21-4
Finestra di dialogo Analisi discriminante: Statistiche



Descrittive. Le opzioni disponibili sono medie (incluse le deviazioni standard), ANOVA univariate e test *M* di Box.

- **Medie.** Visualizza le medie e le deviazioni standard globali e di gruppo delle variabili indipendenti.
- **ANOVA univariate.** Effettua l'analisi della varianza univariata per verificare l'uguaglianza delle medie di gruppo per ciascuna variabile indipendente.
- **M di Box.** Un test per l'uguaglianza di matrici di covarianza di gruppo. Per dimensioni di campione sufficientemente elevate, un valore P non significativo vuol dire che non ci sono sufficienti prove che le matrici differiscano. Il test è sensibile a scostamenti dalla normalità multivariata, tende cioè a non considerare uguali le matrici se l'ipotesi di normalità viene violata.

Coefficienti di funzione. Le opzioni disponibili sono i coefficienti di correlazione di Fisher e i coefficienti non standardizzati.

- **Di Fisher (Analisi discriminante).** Visualizza i coefficienti di Fisher della funzione discriminante, che possono essere usati direttamente per la classificazione. Viene riprodotto un insieme separato di coefficienti di funzioni di classificazione per ciascun gruppo. Ogni caso viene assegnato al gruppo in cui ottiene il più alto punteggio discriminante (valore della funzione di classificazione).
- **Non standardizzati.** Visualizza i coefficienti della funzione discriminante non standardizzati.

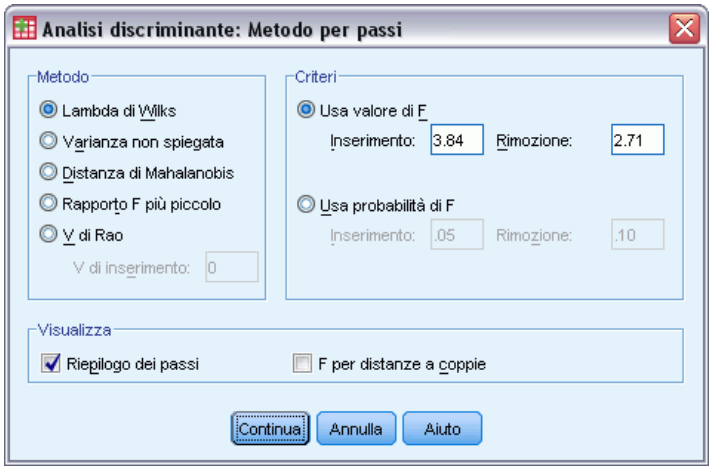
Matrici. Le matrici di coefficienti disponibili per le variabili indipendenti sono: matrice di correlazione entro gruppi, matrice di covarianza entro gruppi, matrice di covarianza gruppi separati e matrice di covarianza totale.

- **Correlazione entro gruppi.** Visualizza la matrice di correlazione entro gruppi ottenuta mediando le matrici di covarianza di tutti i gruppi prima di calcolare le correlazioni.
- **Covarianza entro gruppi.** Visualizza una matrice combinata di covarianza entro i gruppi, mediando le singole matrici di covarianza di tutti i gruppi. Questa matrice può essere diversa dalla matrice di covarianza globale.

- **Covarianza di gruppi separati.** Visualizza matrici di covarianza separate per ciascun gruppo.
- **Covarianza totale.** Visualizza la matrice di covarianza globale, calcolata su tutti i casi, ovvero ignorando la suddivisione in gruppi.

Analisi discriminante: Metodo Stepwise

Figura 21-5
Finestra di dialogo Analisi discriminante: Metodo Stepwise



Metodo. Selezionare la statistica da utilizzare per inserire o rimuovere nuove variabili. Le alternative disponibili sono: lambda di Wilks, varianza non spiegata, distanza di Mahalanobis, minimo rapporto *F* e *V* di Rao. Con il *V* di Rao è possibile specificare l'aumento minimo di *V* per la variabile da inserire.

- **Lambda di Wilks.** Un metodo di selezione delle variabili nell'analisi discriminante per passi che sceglie le variabili da inserire nell'equazione in base a quanto esse contribuiscono a minimizzare il Lambda di Wilks. Ad ogni passo viene inserita la variabile che minimizza il valore globale del Lambda di Wilks'.
- **Varianza non spiegata.** Ad ogni passo viene inserita la variabile che riduce al minimo la somma della variazione spiegata fra gruppi.
- **Distanza di Mahalanobis.** Una misura della distanza di un caso dalla media di tutti i casi per le variabili indipendenti. Un'elevata distanza di Mahalanobis indica che un caso include valori estremi per una o più variabili indipendenti.
- **Rapporto F più piccolo.** Un metodo di selezione delle variabili nelle analisi per passi basato sulla massimizzazione di un rapporto F valutato tramite la distanza di Mahalanobis tra gruppi.
- **V di Rao.** Una misura delle differenze tra medie di gruppo. Detta anche traccia di Lawley-Hotelling. Ad ogni passo viene inserita la variabile che massimizza l'aumento della V di Rao. Specificare l'incremento minimo che una variabile deve apportare per essere inserita nel modello.

Criteri. Le alternative disponibili sono: Usa valore di F e Usa probabilità di F. Immettere valori per aggiungere o rimuovere variabili.

- **Usa valore di F.** La variabile viene inserita nel modello se il relativo valore F è maggiore di quello di inserimento. La variabile viene altresì rimossa se il relativo valore F è minore di quello di rimozione. I valori di inserimento e di rimozione devono essere entrambi positivi e Inserimento deve essere maggiore di Rimozione Abbassando il valore di inserimento e/o alzando quello di rimozione si allentano i vincoli di inclusione delle variabili nel modello.
- **Usa probabilità di F.** La variabile viene inserita nel modello se il livello di significatività del relativo valore di F è minore di quello di inserimento. La variabile viene altresì rimossa se il livello di significatività è maggiore di quello di rimozione. I valori di inserimento e di rimozione devono essere entrambi positivi e Inserimento deve essere minore di Rimozione. Alzando il valore di inserimento e/o abbassando quello di rimozione si allentano i vincoli di inclusione delle variabili nel modello.

Visualizzazione. Riepilogo dei passi visualizza le statistiche per tutte le variabili dopo ogni passo. F per distanze a coppie visualizza una matrice di valori *F* pairwise per ciascuna coppia di gruppi.

Analisi discriminante: Classificazione

Figura 21-6
Finestra di dialogo Analisi discriminante: Classificazione

Probabilità a priori. Questa opzione determina se i coefficienti di classificazione vengono corretti per una conoscenza a priori del gruppo di appartenenza.

- **Tutti i gruppi uguali.** Si presuppongono probabilità a priori uguali per tutti i gruppi, senza effetti sui coefficienti.
- **Calcola dalle dimensioni dei gruppi.** Le dimensioni del gruppo osservate determinano le probabilità a priori del gruppo di appartenenza. Ad esempio, se il 50% delle osservazioni incluse nell'analisi rientrano nel primo gruppo, il 25% nel secondo e il 25% nel terzo, i coefficienti di classificazione vengono corretti in modo da aumentare la probabilità di appartenenza nel primo gruppo rispetto agli altri due.

Visualizzazione. Le opzioni di visualizzazione disponibili sono: risultati per casi, tabella riassuntiva e classificazione autoesclusiva.

- **Risultati per casi.** Visualizza per ciascun caso i codici del gruppo effettivo, del gruppo previsto, della probabilità a posteriori e del punteggio discriminante.

- **Tabella riassuntiva.** Il numero di casi assegnati in modo corretto e non corretto a ciascuno dei gruppi in base all'analisi discriminante. A volte detta "Matrice confusione".
- **Classificazione autoesclusiva.** Ogni caso viene classificato usando le funzioni costruite su tutti i casi meno se stesso. Nota anche come classificazione leave one out o metodo U.

Sostituisci valori mancanti con la media. Selezionare questa opzione per sostituire un valore mancante con la media di una variabile indipendente, solo durante la fase di classificazione.

Usa matrice di covarianza. È possibile classificare i casi utilizzando una matrice di covarianza entro gruppi o una matrice di covarianza gruppi separati.

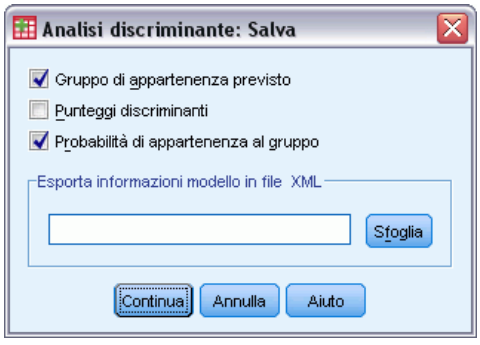
- **Entro gruppi.** Per classificare i casi viene utilizzata la matrice globale di covarianza entro i gruppi.
- **Gruppi separati.** Per classificare i casi vengono utilizzate le matrici di covarianza dei singoli gruppi. Dal momento che la classificazione è basata sulla funzione discriminante e non sui valori originali, questa opzione non è sempre equivalente alla discriminazione quadratica.

Grafici. Le opzioni disponibili per i grafici sono: gruppi accorpati, gruppi separati e mappa territoriale.

- **Gruppi accorpati.** Produce un unico grafico a dispersione dei valori delle prime due funzioni discriminanti. Se c'è una sola funzione, viene prodotto un istogramma.
- **Gruppi separati.** Produce un grafico a dispersione dei valori delle prime due funzioni discriminanti per ciascun gruppo. Se c'è una sola funzione, verranno prodotti istogrammi.
- **Mappa territoriale.** Un grafico dei confini usati per classificare i casi in gruppi in base ai valori di una funzione. I numeri corrispondono ai gruppi nei quali vengono classificati i casi. La media per ciascun gruppo è indicata da un asterisco all'interno dei suoi confini. La mappa non viene visualizzata se c'è una sola funzione discriminante.

Analisi discriminante: Salva

Figura 21-7
Finestra di dialogo Analisi discriminante: Salva



È possibile aggiungere nuove variabili al file di dati attivo. Le opzioni disponibili sono: gruppo di appartenenza previsto (una sola variabile), punteggi discriminanti (una variabile per ciascuna funzione discriminante nella soluzione) e probabilità di appartenenza al gruppo in base ai punteggi discriminanti (una variabile per ciascun gruppo).

È possibile esportare le informazioni sul modello nel file specificato in formato XML (PMML). È possibile utilizzare questo file di modello per applicare le informazioni del modello ad altri file di dati per il calcolo del punteggio.

Funzioni aggiuntive del comando DISCRIMINANT

Il linguaggio della sintassi dei comandi consente inoltre di:

- Eseguire più analisi discriminanti (con un comando) e controllare l'ordine in cui le variabili vengono inserite (con il sottocomando `ANALYSIS`).
- Specificare le probabilità a priori per la classificazione (con il sottocomando `PRIORS`).
- Visualizzare le matrici dei modelli e delle strutture ruotate (con il sottocomando `ROTATE`).
- Limitare il numero di funzioni discriminanti estratte (con il sottocomando `FUNCTIONS`).
- Limitare la classificazione ai casi selezionati (o non selezionati) per l'analisi (con il sottocomando `SELECT`).
- Leggere e analizzare la matrice di correlazione (con il sottocomando `MATRIX`).
- Scrivere una matrice di correlazione da analizzare in seguito (con il sottocomando `MATRIX`).

Vedere *Command Syntax Reference* per informazioni dettagliate sulla sintassi.

Capitolo
22

Analisi fattoriale

L'analisi fattoriale si propone di identificare le variabili sottostanti, o **fattori**, che illustrano il modello per le correlazioni all'interno di un insieme di variabili osservate. L'analisi fattoriale viene in genere utilizzata per la riduzione dei dati in quanto consente di identificare un numero ridotto di valori che spiegano la maggior parte dei valori di varianza osservati in numerose variabili manifeste. L'analisi fattoriale può inoltre essere utilizzata per generare ipotesi relative a meccanismi causali oppure per esaminare le variabili per le analisi successive (ad esempio per identificare la collinearità prima di eseguire un'analisi di regressione lineare).

La procedura di analisi fattoriale permette un elevato grado di flessibilità:

- Sono disponibili sette metodi di estrazione fattoriale.
- Sono disponibili cinque metodi di rotazione, tra cui oblimin diretto e promax per le rotazioni non ortogonali.
- Sono disponibili tre metodi per il calcolo dei punteggi, che possono essere salvati come variabili per le analisi successive.

Esempio. Quali sono gli atteggiamenti sottostanti che inducono le persone a rispondere a questionari politici in un determinato modo? Dall'esame delle correlazioni esistenti tra le voci del questionario risulta una significativa sovrapposizione tra diversi sottogruppi di voci. Ad esempio, le domande relative alle tasse tendono ad essere correlate fra loro, così come le domande relative alle questioni militari e così via. Grazie all'analisi fattoriale è possibile identificare il numero di fattori sottostanti e in molti casi determinare cosa rappresentano concettualmente tali fattori. È inoltre possibile calcolare i punteggi fattoriali per ciascun rispondente, un elemento che è possibile utilizzare in analisi successive. Ad esempio, è possibile creare un modello di regressione logistico che consenta di prevedere il comportamento di voto in base ai punteggi fattoriali.

Statistiche. Per ogni variabile: numero di casi validi, media e deviazione standard. Per ciascuna analisi fattoriale: matrice di correlazione delle variabili, inclusi i livelli di significatività, determinante, inversa; matrice di correlazione riprodotta, inclusa anti-immagine; soluzione iniziale (comunalità, autovalori e percentuale di varianza spiegata); misura di adeguatezza campionaria di Kaiser-Meyer-Olkin e test di sfericità di Bartlett; soluzione non ruotata, inclusi pesi fattoriali, comunalità e autovalori; soluzione ruotata, incluse la matrice ruotata dei modelli e la matrice di trasformazione. Per le rotazioni oblique: matrice ruotata dei modelli e delle strutture; matrice dei coefficienti di punteggio fattoriale e matrice di covarianza fattoriale. Grafici: grafico decrescente degli autovalori e grafico dei pesi fattoriali dei primi due o tre fattori.

Dati. Le variabili devono essere quantitative al livello di misura per **intervallo** o per **rapporto**. I dati categoriali (ad esempio religione o paese d'origine) non sono idonei per l'analisi fattoriale. I dati per cui è possibile calcolare i coefficienti di correlazione di Pearson sono idonei per l'analisi fattoriale.

Assunzioni. I dati devono avere una distribuzione normale bivariata per ciascuna coppia di variabili e le osservazioni devono essere indipendenti. Il modello di analisi fattoriale specifica che le variabili vengono determinate da fattori comuni (i fattori stimati dal modello) e fattori univoci

(che non risultano sovrapposti tra le variabili osservate); le stime calcolate si basano sull'ipotesi che tutti i fattori univoci siano correlati reciprocamente e con i fattori comuni.

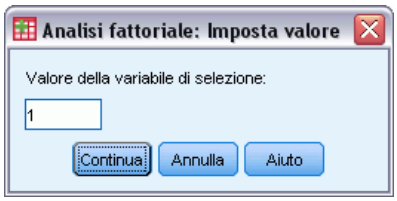
Per ottenere un'analisi fattoriale

- Dai menu, scegliere:
  Analizza > Riduzioni dimensione > Fattoriale...
- Selezionare le variabili per l'analisi fattoriale.

Figura 22-1
Finestra di dialogo Analisi fattoriale

Analisi fattoriale: Seleziona casi

Figura 22-2
Finestra di dialogo Analisi fattoriale: Imposta valore



Per selezionare i casi da usare nell'analisi:

- Selezionare una variabile di selezione.
- Fare clic suValore per immettere la variabile di selezione come numero intero.

  Nell'analisi fattoriale verranno utilizzati solo i casi con tale valore per la variabile di selezione.

Analisi fattoriale: Descrittive

Figura 22-3
Finestra di dialogo Analisi fattoriale: Descrittive

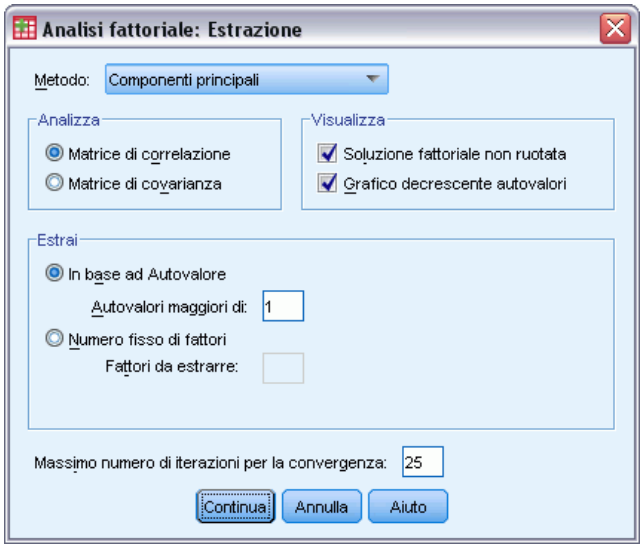
Statistiche. Le statistiche descrittive univariate includono la media, la deviazione standard e il numero di casi validi per ciascuna variabile. Nella soluzione iniziale vengono visualizzate le comunalità iniziali, gli autovalori e la percentuale della varianza spiegata.

Matrice di correlazione. Le opzioni disponibili sono: coefficienti, livelli di significatività, determinante, inversa, riprodotta, anti-immagine, test KMO e test di sfericità di Bartlett.

- **Test KMO e test di sfericità di Bartlett.** La misura di adeguatezza campionaria KMO (Keiser Meyer Olkin) verifica se le correlazioni parziali tra le variabili sono piccole. Il test di sfericità di Bartlett verifica se la matrice di correlazione è una matrice identità, cosa che indicherebbe l'inadeguatezza del modello fattoriale.
- **Riprodotta.** La matrice di correlazione stimata a partire dalla soluzione del fattore. Vengono visualizzati anche i residui (differenze tra correlazioni stimate e osservate).
- **Anti-immagine.** La matrice di correlazione anti-immagine contiene i coefficienti di correlazione parziale cambiati di segno e la matrice di covarianza anti-immagine contiene le covarianze parziali cambiate di segno. In un buon modello fattoriale, la maggior parte degli elementi fuori dalla diagonale dovrebbe avere valori bassi. La misura di adeguatezza campionaria di una variabile è visualizzata sulla diagonale della matrice di correlazione anti-immagine.

Analisi fattoriale: Estrazione

Figura 22-4
Finestra di dialogo Analisi fattoriale: Estrazione



Metodo. Consente di specificare il metodo di estrazione dei fattori. I metodi disponibili sono: componenti principali, minimi quadrati non ponderati, minimi quadrati generalizzati, massima verosimiglianza, fattorizzazione dell'asse principale, fattorizzazione alfa e fattorizzazione immagine.

- **Componenti principali (Analisi fattoriale).** Metodo usato per formare combinazioni lineari non correlate delle variabili osservate. La prima componente spiega la parte più alta di variabilità. Le componenti successive spiegano porzioni di variabilità decrescenti e sono tutte non correlate fra loro. L'analisi delle componenti principali viene usata per ottenere la soluzione fattoriale iniziale. Può essere usata quando una matrice di correlazione è singolare.
- **Metodo dei minimi quadrati non ponderati.** Un metodo per l'estrazione dei fattori che minimizza la somma dei quadrati delle differenze tra la matrice di correlazione osservata e quella riprodotta, ignorando le diagonali.
- **Minimi quadrati generalizzati (Analisi fattoriale).** Un metodo di estrazione dei fattori che minimizza la somma dei quadrati delle differenze tra la matrice di correlazione osservata e la matrice di correlazione riprodotta. Le correlazioni sono ponderate tramite l'inverso della loro unicità, in modo da dare meno peso alle variabili con elevata unicità rispetto a quelle con unicità inferiore.
- **Massima verosimiglianza (Analisi fattoriale).** Un metodo per l'estrazione dei fattori che produce le stime dei parametri che più verosimilmente hanno prodotto la matrice di correlazione osservata, se il campione è estratto da una distribuzione normale multivariata. Le correlazioni sono pesate tramite l'inverso dell'unicità delle variabili. Viene utilizzato un algoritmo iterativo.
- **Fattorizzazione dell'asse principale.** Un metodo di estrazione dei fattori dalla matrice di correlazione originale con coefficienti di correlazione multipla al quadrato posti sulla diagonale come stime iniziali delle comunalità. Questi pesi di fattore vengono usati per stimare nuove comunalità che sostituiscono le vecchie stime sulla diagonale. Le iterazioni

continuano fino a che le variazioni nelle comunalità da un'iterazione alla successiva soddisfano il criterio di convergenza per l'estrazione.

- **Alfa.** Un metodo per l'estrazione dei fattori che considera le variabili nell'analisi come un campione estratto dall'universo delle variabili potenziali. Questo metodo massimizza l'Alpha di Cronbach dei fattori.
- **Fattorizzazione immagine (Analisi fattoriale).** Metodo di estrazione dei fattori sviluppato da Guttman e basato sulla teoria dell'immagine. La parte comune della variabile, detta immagine parziale, viene definita come la sua regressione lineare sulle variabili rimanenti, piuttosto che in funzione di fattori ipotetici.

Analizza. Consente di specificare una matrice di correlazione o una matrice di covarianza.

- **Matrice di correlazione.** Può risultare utile se le variabili dell'analisi vengono misurate su scale diverse.
- **Matrice di covarianza.** Può risultare utile se si intende applicare l'analisi fattoriale a più gruppi con varianze diverse per ogni variabile.

Estrai. È possibile mantenere tutti i fattori con autovalori superiori al valore specificato oppure mantenere solo il numero di fattori specificato.

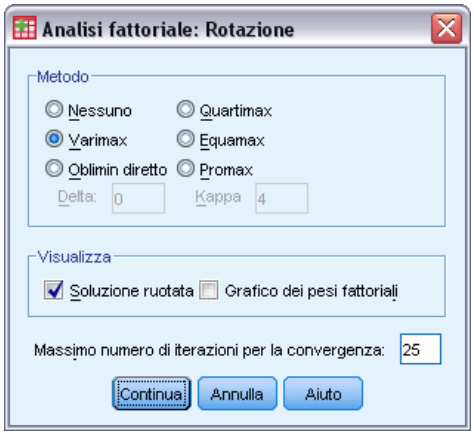
Visualizzazione. Consente di richiedere la soluzione fattoriale non ruotata e un grafico decrescente degli autovalori.

- **Soluzione non ruotata.** Visualizza la matrice dei pesi fattoriali, le comunalità e gli autovalori della soluzione fattoriale.
- **Grafico decrescente autovalori.** Un grafico della varianza associata a ciascun fattore. Questo grafico viene usato per determinare il numero di fattori da mantenere. In genere il grafico mostra una decisa diminuzione di pendenza nel punto in cui entrano in gioco i fattori meno rilevanti.

Massimo numero di iterazioni per la convergenza. Consente di specificare il numero massimo di passaggi che l'algoritmo può eseguire per valutare la soluzione.

Analisi fattoriale: Rotazione

Figura 22-5
Finestra di dialogo Analisi fattoriale: Rotazione



Metodo. Consente di selezionare il metodo di rotazione fattoriale. I metodi disponibili sono: varimax, equamax, quartimax, oblimin diretto e promax.

- **Varimax (Analisi fattoriale).** Un metodo di rotazione ortogonale che minimizza il numero di variabili che caricano fortemente ciascun fattore. Questo metodo semplifica l'interpretazione dei fattori.
- **Metodo oblimin diretto.** Un metodo di rotazione obliqua (non ortogonale). Quando delta vale 0 (il valore di default), le soluzioni sono per la maggior parte oblique. Quando delta diventa negativo e aumenta in valore assoluto, i fattori cominciano a essere meno obliqui. Inserire un numero minore o uguale a 0,8 per sovrascrivere il valore di default.
- **Metodo Quartimax.** Un metodo di rotazione che rende minimo il numero di fattori necessari per spiegare ogni variabile. Questo metodo semplifica l'interpretazione delle variabili osservate.
- **Equamax.** Un metodo di rotazione che rappresenta una combinazione tra il metodo varimax, che minimizza i fattori, e il metodo quartimax, che semplifica le variabili. È una combinazione dei metodi Varimax e Quartimax.
- **Rotazione Promax.** Una rotazione obliqua che ammette la correlazione fra fattori. Questa rotazione può essere utile per file di grandi dimensioni in quanto è più veloce del metodo Oblimin.

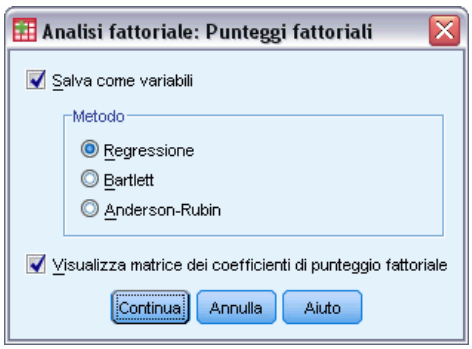
Visualizzazione. Consente di includere l'output nella soluzione ruotata, nonché i grafici dei pesi fattoriali per i primi due o tre fattori.

- **Soluzione ruotata (Analisi fattoriale).** Per ottenere la soluzione ruotata occorre aver selezionato un metodo di rotazione. Per le rotazioni ortogonali vengono visualizzate la matrice ruotata dei modelli e la matrice di trasformazione. Per le rotazioni oblique vengono visualizzate la matrice dei modelli, la matrice di struttura e la matrice di correlazione dei fattori.
- **Grafico dei pesi fattoriali.** Grafico tridimensionale dei pesi dei primi tre fattori. Per le soluzioni a due fattori viene prodotto un grafico bidimensionale. Assente se l'analisi estrae un solo fattore. Se è stata richiesta la rotazione, il grafico visualizza la soluzione ruotata.

Massimo numero di iterazioni per la convergenza. Consente di specificare il massimo numero di passaggi che l'algoritmo può eseguire per completare la rotazione.

Analisi fattoriale: Punteggi fattoriali

Figura 22-6
Finestra di dialogo Analisi fattoriale: Punteggi fattoriali



Salva come variabili. Consente di creare una nuova variabile per ciascun fattore nella soluzione finale.

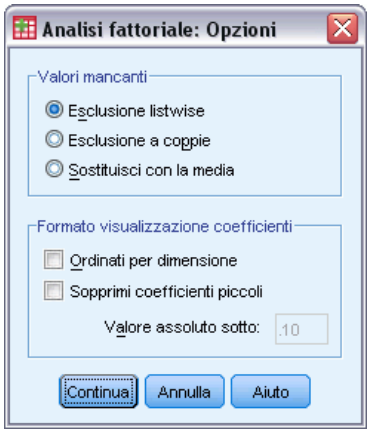
Metodo. I metodi alternativi per il calcolo dei punteggi fattoriali sono regressione, Bartlett e Anderson-Rubin.

- **Metodo di regressione.** Un metodo per calcolare i coefficienti dei punteggi fattoriali. I punteggi prodotti hanno media 0 e varianza pari al quadrato della correlazione multipla fra i punteggi stimati e i valori reali dei fattori. I punteggi possono essere correlati anche quando i fattori sono ortogonali.
- **Punteggi di Bartlett.** Un metodo di stima dei coefficienti di punteggio fattoriale. I punteggi fattoriali di Bartlett hanno media pari a 0. La somma dei quadrati dei singoli fattori sull'intervallo delle variabili è minimizzata.
- **Metodo di Anderson-Rubin.** Un metodo per calcolare i coefficienti dei punteggi fattoriali; rappresenta una modifica del metodo di Bartlett per assicurare l'ortogonalità dei fattori stimati. I punteggi forniti hanno una media pari a 0, deviazione standard pari a 1 e risultano ortogonali (non correlati fra loro).

Visualizza matrice dei coefficienti di punteggio fattoriale. Mostra i coefficienti per cui vengono moltiplicate le variabili per ottenere i punteggi fattoriali. Vengono visualizzate anche le correlazioni tra i punteggi fattoriali.

Analisi fattoriale: Opzioni

Figura 22-7
Finestra di dialogo Analisi fattoriale: Opzioni



Valori mancanti. Consente di specificare le modalità di gestione dei valori mancanti. Le scelte disponibili sono: Esclusione **listwise**, Esclusione **pairwise** o Sostituisci con la media.

Formato visualizzazione coefficienti. Consente di controllare alcuni aspetti delle matrici di output. È possibile ordinare i coefficienti per dimensioni ed eliminare i coefficienti con valori assoluti inferiori al valore specificato.

Opzioni aggiuntive del comando FACTOR

Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare i criteri di convergenza per l'iterazione durante l'estrazione e la rotazione.
- Specificare i singoli grafici dei fattori ruotati.
- Specificare quanti punteggi di fattori salvare.
- Specificare i valori diagonali per il metodo di calcolo dei fattori dell'asse principale.
- Scrivere le matrici di correlazione o le matrici dei fattori sul disco per poterle analizzare in seguito.
- Leggere e analizzare le matrici di correlazione o dei fattori.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
23

Scelta di una procedura per il raggruppamento

È possibile eseguire cluster analysis utilizzando le procedure Analisi Cluster TwoStep, Cluster gerarchica o Cluster con metodo delle k-medie. Ogni procedura utilizza un algoritmo diverso per la creazione di cluster e include opzioni non disponibili nelle altre procedure.

Analisi Cluster TwoStep. La procedura Analisi Cluster TwoStep risulta appropriata per molte applicazioni. Rende disponibili le seguenti funzioni univoche:

- Selezione automatica del numero di cluster più appropriato, oltre a misure per la scelta tra modelli di cluster.
- Possibilità di creare contemporaneamente due modelli di cluster basati su variabili categoriali e continue.
- Possibilità di salvare il modello di cluster in un file XML esterno e quindi di leggere tale file e aggiornare il modello di cluster utilizzando i dati più recenti.

Inoltre, questa procedura consente di analizzare file di dati di grandi dimensioni.

Cluster gerarchica. La procedura Cluster gerarchica è limitata a file di dati di dimensioni minori (ad esempio per il raggruppamento di centinaia di oggetti) e rende disponibili le seguenti funzioni univoche:

- Possibilità di raggruppare casi o variabili in cluster.
- Possibilità di calcolare un intervallo di soluzioni possibili e di salvare l'appartenenza a un cluster per ognuna di queste soluzioni.
- Diversi metodi per la formazione di cluster, la trasformazione delle variabili e la misurazione della dissimilarità tra cluster.

La procedura Cluster gerarchica analizza variabili per intervallo (continue), binarie o di conteggio, purché le variabili siano tutte dello stesso tipo.

Cluster con metodo delle k-medie. La procedura Cluster con metodo delle k-medie è limitata a dati continui e richiede che il numero di cluster venga specificato anticipatamente, rendendo tuttavia disponibili le seguenti funzioni univoche:

- Possibilità di salvare le distanze dai centri di cluster per ogni oggetto.
- Possibilità di leggere centri iniziali di cluster da un file SPSS esterno e salvare centri finali di cluster in un file esterno in formato IBM® SPSS® Statistics.

Inoltre, questa procedura consente di analizzare file di dati di grandi dimensioni.

Capitolo
24

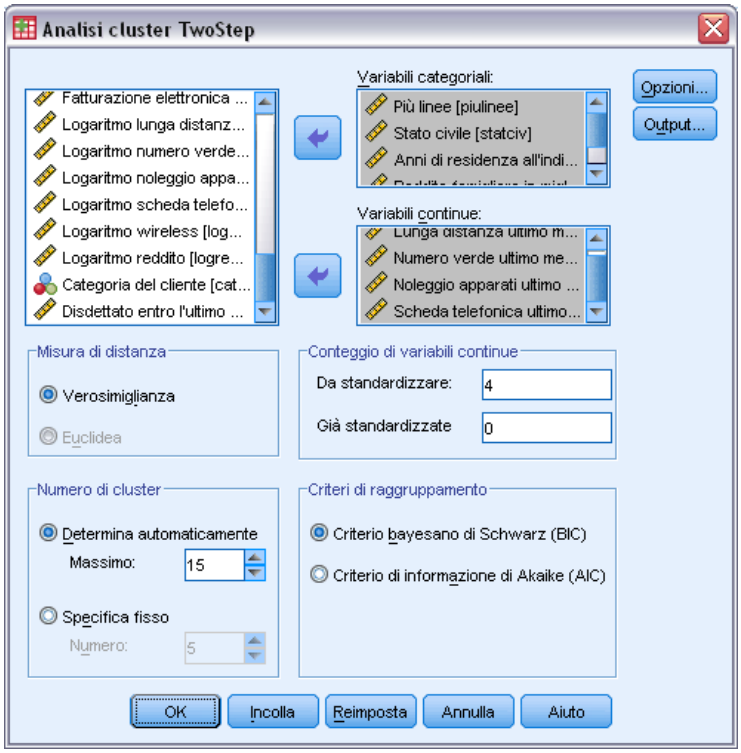
Analisi cluster TwoStep

L'analisi cluster TwoStep è uno strumento di esplorazione che consente di rilevare raggruppamenti naturali, o cluster, all'interno di insiemi di dati, che non sarebbero altrimenti evidenti. L'algoritmo utilizzato da questa procedura presenta diverse caratteristiche che lo differenziano dalle tecniche di raggruppamento tradizionali:

- **Gestione di variabili categoriali e continue.** Se le variabili sono indipendenti, è possibile applicare una distribuzione normale multinomiale congiunta alle variabili categoriali e continue.
- **Selezione automatica del numero di cluster.** Mediante il confronto tra i valori dei criteri di scelta di modello appartenenti a diverse soluzioni di raggruppamento, la procedura è in grado di determinare automaticamente il numero ottimale di cluster.
- **Scalabilità.** Mediante la creazione di un albero delle caratteristiche dei cluster (CF) che fornisce un riepilogo dei record, l'algoritmo TwoStep consente di analizzare file di dati di grandi dimensioni.

Esempio. I produttori di articoli al dettaglio e prodotti per i consumatori applicano regolarmente tecniche di raggruppamento ai dati relativi alle abitudini di acquisto dei propri clienti, al sesso, all'età, al livello di reddito e così via. In questo modo adattano le strategie di sviluppo del prodotto e di mercato ad ogni gruppo di consumatori al fine di aumentare le vendite e accrescere la fedeltà alla marca.

Figura 24-1
Finestra di dialogo Analisi cluster TwoStep



Misura di distanza. Questa selezione determina la modalità di calcolo della similarità tra due cluster.

- **Verosimiglianza.** La misura di verosimiglianza applica una distribuzione per probabilità alle variabili. Si suppone che le variabili continue vengano distribuite normalmente, mentre le variabili categoriali in base al modello multinomiale. Si suppone che tutte le variabili siano indipendenti.
- **Euclidea.** La misura euclidea è la distanza in "linea retta" tra due cluster. Può essere utilizzata solo quando tutte le variabili sono continue.

Numero di cluster. Questa selezione consente di specificare la modalità di definizione del numero dei cluster.

- **Determina automaticamente.** La procedura determina automaticamente il numero di cluster ottimale, mediante i criteri specificati nel gruppo Criteri di raggruppamento. È inoltre possibile immettere un numero intero positivo per definire il numero di cluster massimo che la procedura dovrà prendere in considerazione.
- **Specifica fisso.** Consente di definire un numero fisso di cluster nella soluzione. Immettere un numero intero positivo.

Conteggio di variabili continue. Questo gruppo fornisce un riepilogo delle opzioni di standardizzazione relative alle variabili continue specificate nella finestra di dialogo Opzioni. Per ulteriori informazioni, vedere l'argomento Opzioni di Analisi cluster TwoStep a pag. 170.

Criteri di raggruppamento. Questa selezione determina la modalità di definizione del numero dei cluster mediante l'algoritmo di raggruppamento automatico. È possibile specificare il modello Criterio bayesiano di Schwarz (BIC, Bayesian Information Criterion) o il modello Criterio di informazione di Akaike (AIC, Akaike Information Criterion).

Dati. Questa procedura può essere utilizzata sia con le variabili continue sia con le variabili categoriali. I casi rappresentano gli oggetti da raggruppare e le variabili corrispondono agli attributi in base ai quali viene eseguito il raggruppamento.

Ordine dei casi. La funzione Albero delle caratteristiche cluster e la soluzione finale possono dipendere dall'ordine dei casi. Per ridurre al minimo gli effetti dell'ordine, disporre i casi in ordine casuale. Può essere utile ottenere più soluzioni diverse con casi disposti in ordini casuali diversi per verificare la stabilità di una soluzione specifica. Nei casi in cui questa operazione è complessa a causa delle dimensioni eccessive dei file, è possibile effettuare più operazioni con un campione di casi disposti in ordini casuali diversi.

Assunzioni. La misura di distanza della verosimiglianza assume che le variabili nel modello di cluster siano indipendenti. A ogni variabile continua inoltre si suppone inoltre che sia associata una distribuzione normale o gaussiana, mentre a ogni variabile categoriale una distribuzione multinomiale. La verifica empirica interna indica che la procedura è piuttosto robusta rispetto alle violazioni sia delle assunzioni di indipendenza sia delle assunzioni di distribuzione, ma è consigliabile verificare fino a che punto tali assunzioni vengano soddisfatte.

Utilizzare la procedura Correlazioni bivariate per verificare l'indipendenza di due variabili continue. Utilizzare la procedura Tavole di contingenza per verificare l'indipendenza di due variabili categoriali. Utilizzare la procedura Medie per verificare l'indipendenza tra una variabile continua e una variabile categoriale. Utilizzare la procedura Esplora per verificare la normalità di una variabile continua. Utilizzare la procedura Test Chi-quadrato per verificare se per una variabile categoriale è stata specificata una distribuzione multinomiale.

Per ottenere un'analisi cluster TwoStep

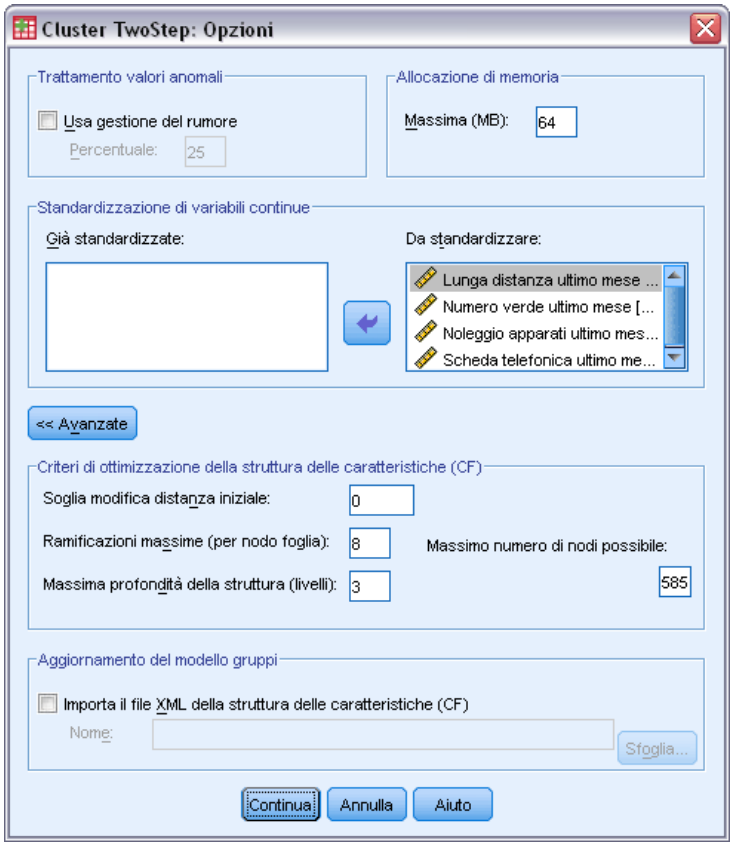
- Dai menu, scegliere:
  Analizza > Classifica > Cluster TwoStep...
- Selezionare una o più variabili categoriali o continue.

Se lo si desidera, è possibile:

- Modificare i criteri in base ai quali sono stati creati i cluster.
- Selezionare le impostazioni di gestione del rumore, allocazione di memoria, standardizzazione delle variabili e input del modello di cluster.
- Richiedere l'output in Viewer modelli.
- Salvare i risultati del modello nel file di lavoro o in un file XML esterno.

Opzioni di Analisi cluster TwoStep

Figura 24-2
Finestra di dialogo Cluster TwoStep: Opzioni



Trattamento valori anomali. Questo gruppo consente di trattare i valori anomali soprattutto durante il raggruppamento, se l'albero delle caratteristiche dei cluster risulta pieno. L'albero delle caratteristiche dei cluster è pieno quando non è più in grado di accettare casi in un nodo foglia e nessun nodo foglia può essere suddiviso.

- Se si seleziona la gestione del rumore e l'albero delle caratteristiche dei cluster risulta pieno, sarà possibile ampliarlo posizionando i casi mal distribuiti in più foglie all'interno di una foglia specifica per il "rumore". Una foglia contiene casi mal distribuiti quando il numero dei casi è inferiore alla percentuale specificata per la dimensione massima della foglia. Dopo aver ampliato la struttura, i valori anomali vengono inseriti nell'albero delle caratteristiche dei cluster, se possibile. Altrimenti, vengono eliminati.
- Se non si seleziona la gestione del rumore e l'albero delle caratteristiche dei cluster risulta pieno, sarà possibile ampliarlo utilizzando una soglia di modifica della distanza più elevata. Dopo il raggruppamento finale, i valori non assegnati a un cluster vengono definiti valori anomali. Al cluster dei valori anomali viene assegnato il numero di identificazione –1 e non viene incluso nel conteggio dei cluster.

Allocazione di memoria. Questo gruppo consente di specificare in megabyte (MB) la quantità massima di memoria che l'algoritmo del cluster può utilizzare. Se questa quantità massima viene superata, la procedura utilizzerà il disco per memorizzare le informazioni che non è possibile inserire nella memoria. Specificare un numero maggiore o uguale a 4.

- Per informazioni sul valore massimo per il sistema, rivolgersi all'amministratore del sistema.
- La ricerca dei cluster corretti o desiderati mediante l'algoritmo potrebbe non riuscire, se il valore è troppo basso.

Standardizzazione delle variabili. L'algoritmo di raggruppamento funziona con variabili continue standardizzate. Qualsiasi variabile non standardizzata deve essere impostata come "Da standardizzare". Per risparmiare tempo e calcoli, è possibile impostare le variabili continue già standardizzate come "Già standardizzate".

Opzioni avanzate

Criteri di ottimizzazione dell'albero delle caratteristiche (CF). Le seguenti impostazioni dell'algoritmo di raggruppamento riguardano in modo specifico l'albero delle caratteristiche dei cluster e devono essere modificate con cautela:

- **Soglia modifica distanza iniziale.** Si tratta della soglia iniziale utilizzata per ampliare l'albero delle caratteristiche dei cluster. Se dopo l'inserimento di un determinato caso in una foglia dell'albero delle caratteristiche dei cluster, la distanza risulta inferiore alla soglia, la foglia non viene suddivisa. Se la distanza supera la soglia, la foglia può essere suddivisa.
- **Ramificazioni massime (per nodo foglia).** Il numero massimo di nodi figlio per un nodo foglia.
- **Massima profondità struttura (livelli).** Il numero massimo di livelli dell'albero delle caratteristiche dei cluster.
- **Massimo numero di nodi possibile.** Indica il numero massimo dei nodi dell'albero delle caratteristiche dei cluster che può essere potenzialmente generato dalla procedura, in base alla funzione $(b^{d+1} - 1) / (b - 1)$, dove b sono le ramificazioni massime e d la profondità massima. Tenere presente che un albero delle caratteristiche dei cluster di dimensioni eccessive può rappresentare un peso considerevole per le risorse del sistema e compromettere quindi le prestazioni della procedura. Ogni nodo richiede un minimo di 16 byte.

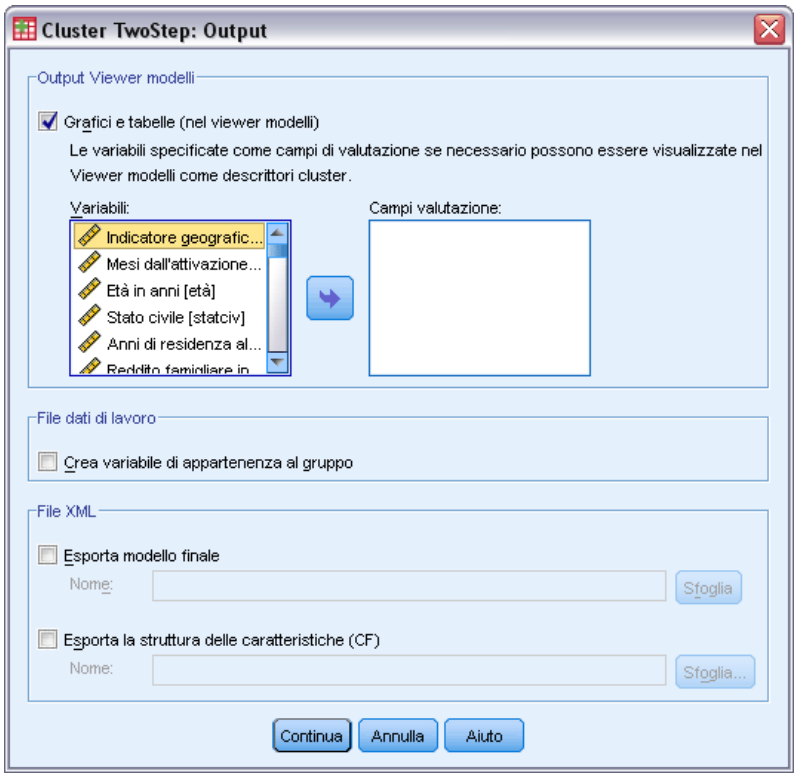
Aggiornamento del modello gruppi. Questo gruppo consente di importare e aggiornare un modello di cluster generato da un'analisi precedente. Il file di input contiene l'albero delle caratteristiche dei cluster in formato XML. Il modello viene quindi aggiornato con i dati nel file attivo. È necessario selezionare i nomi delle variabili nella finestra di dialogo principale in base allo stesso ordine dell'analisi precedente. Il file XML non viene modificato, a meno che le nuove informazioni relative al modello non vengano inserite nello stesso file. Per ulteriori informazioni, vedere l'argomento Output di Analisi cluster TwoStep a pag. 172.

Se viene selezionato l'aggiornamento del modello dei cluster, verranno utilizzate le opzioni per la generazione dell'albero delle caratteristiche dei cluster specificate per il modello originale. Vengono quindi utilizzate le impostazioni del modello salvato relative a misura di distanza, gestione del rumore, allocazione di memoria e ottimizzazione dell'albero delle caratteristiche dei cluster, mentre qualsiasi nuova impostazione specificata nelle finestre di dialogo viene ignorata.

Nota: durante l'aggiornamento di un modello di cluster, la procedura si basa sul presupposto che nessuno dei casi selezionati nel file dati attivo sia stato utilizzato per creare il modello di cluster originale. La procedura si basa inoltre sul presupposto che i casi utilizzati nell'aggiornamento del modello di cluster provengono dalla stessa popolazione di casi utilizzati per creare il modello originale, le medie e le varianze delle variabili continue e i livelli delle variabili categoriali devono quindi essere le stesse per i due insiemi di casi. Se gli insiemi di casi precedenti e correnti provengono da una popolazione eterogenea, è necessario eseguire la procedura Analisi cluster TwoStep negli insiemi di casi combinati per ottenere risultati ottimali.

Output di Analisi cluster TwoStep

Figura 24-3
Finestra di dialogo Cluster TwoStep: Output



Output Viewer modelli. Questo gruppo fornisce le opzioni per la visualizzazione dei risultati dei raggruppamenti.

- **Grafici e tabelle.** Visualizza l'output relativo al modello, tra cui tabelle e grafici. Le tabelle nella visualizzazione tabelle comprendono un riepilogo del modello e una griglia cluster-per-funzioni. L'output grafico nella vista del modello comprende un grafico sulla

qualità dei cluster, le dimensioni dei cluster, l'importanza delle variabili, un grafico di confronto tra cluster e le informazioni sulle celle.

- **Campi di valutazione.** Calcola i dati dei cluster per le variabili che non sono stati utilizzati nella creazione dei cluster. I campi di valutazione possono essere visualizzati insieme alle funzioni di input nel Viewer modelli selezionandoli nella finestra di dialogo secondaria Visualizza. I campi con valori mancanti vengono ignorati.

File dati di lavoro. Questo gruppo consente di salvare le variabili all'interno del file dati attivo.

- **Crea variabile di appartenenza al gruppo.** Questa variabile contiene un numero di identificazione del cluster per ogni caso. Il nome di questa variabile è *tsc_n*, dove *n* è un numero intero positivo che indica l'ordinale dell'operazione di salvataggio del file dati attivo eseguita mediante questa procedura in una determinata sessione.

File XML. Il modello di cluster finale e l'albero delle caratteristiche dei cluster rappresentano due tipi di file di output che possono essere esportati in formato XML.

- **Esporta modello finale.** Il modello di cluster finale viene esportato nel file specificato in formato XML (PMML). È possibile utilizzare questo file di modello per applicare le informazioni del modello ad altri file di dati per il calcolo del punteggio.
- **Esporta l'albero delle caratteristiche (CF).** Questa opzione consente di salvare lo stato corrente dell'albero delle caratteristiche dei cluster e aggiornarlo in un secondo momento utilizzando dati più aggiornati.

Il Visualizzatore cluster

I modelli di cluster vengono solitamente utilizzati per cercare gruppi (o cluster) di record simili in base alle variabili esaminate, in cui la similarità tra i membri dello stesso gruppo è elevata mentre quella tra i membri di gruppi diversi è bassa. È possibile utilizzare i risultati per identificare quelle associazioni che altrimenti non sarebbero evidenti. Ad esempio, attraverso l'analisi dei cluster delle preferenze, del livello di reddito e delle abitudini di spesa dei clienti, è possibile identificare quei tipi di consumatori che con maggiore probabilità risponderanno favorevolmente a una determinata campagna di marketing.

Sono possibili due approcci all'interpretazione dei risultati in una visualizzazione cluster:

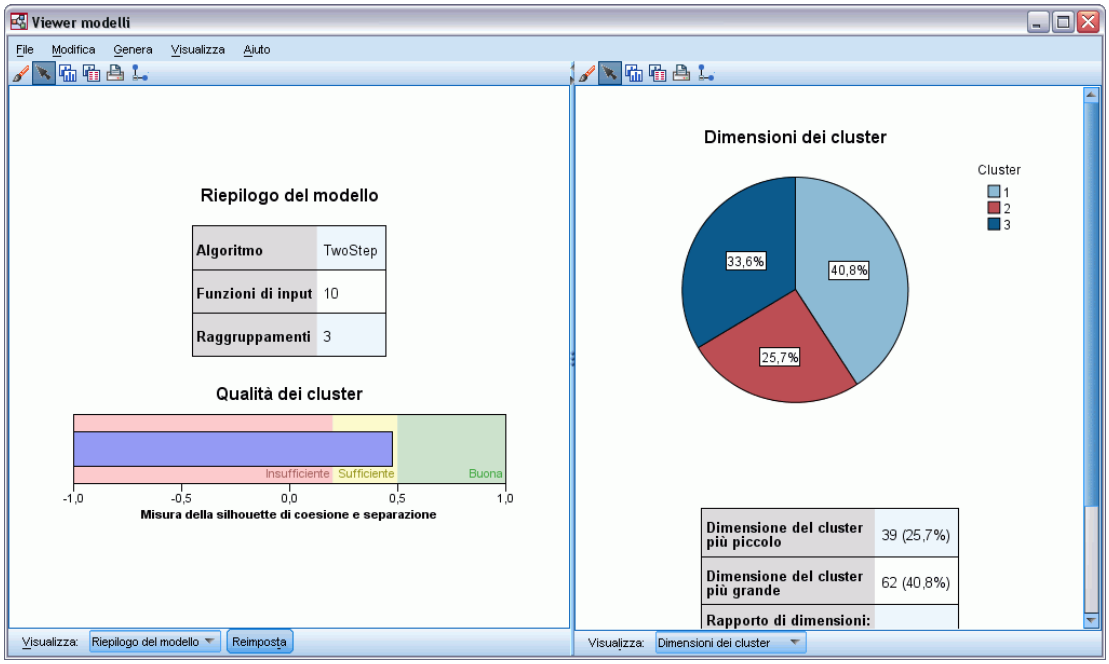
- L'esame di un cluster per determinarne le caratteristiche esclusive. *Il cluster contiene i titolari di prestiti con reddito elevato? Il cluster contiene più record degli altri?*
- L'esame dei campi di tutti i cluster per determinare il modo in cui i valori sono distribuiti tra i cluster. *Il livello di istruzione di un soggetto ne determina l'appartenenza a un cluster? Un punteggio di credito elevato determina l'appartenenza a un cluster o a un altro?*

Mediante le visualizzazioni principali e le diverse visualizzazioni collegate nel Visualizzatore cluster, è possibile ottenere maggiori informazioni per rispondere a questi quesiti.

Per visualizzare le informazioni sul modello di cluster, attivare mediante doppio clic l'oggetto Viewer modelli nel Visualizzatore.

Visualizzatore cluster

Figura 24-4
Schermata predefinita del Visualizzatore cluster



Il Visualizzatore cluster è composto da due riquadri, la visualizzazione principale a sinistra e quella collegata, o ausiliaria, a destra. Le visualizzazioni principali sono due:

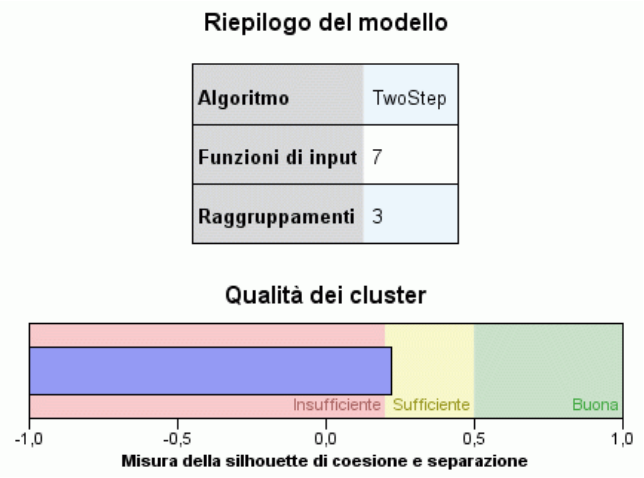
- Riepilogo del modello (visualizzazione predefinita). Per ulteriori informazioni, vedere l'argomento Visualizzazione Riepilogo del modello a pag. 175.
- Raggruppamenti. Per ulteriori informazioni, vedere l'argomento Visualizzazione cluster a pag. 176.

Le visualizzazioni collegate/ausiliarie sono quattro:

- Importanza predittore. Per ulteriori informazioni, vedere l'argomento Visualizzazione Importanza predittore nei cluster a pag. 179.
- Dimensioni cluster (visualizzazione predefinita). Per ulteriori informazioni, vedere l'argomento Visualizzazione Dimensioni dei cluster a pag. 180.
- Distribuzione delle celle. Per ulteriori informazioni, vedere l'argomento Visualizzazione Distribuzione delle celle a pag. 181.
- Confronto tra cluster. Per ulteriori informazioni, vedere l'argomento Visualizzazione Confronto tra cluster a pag. 182.

Visualizzazione Riepilogo del modello

Figura 24-5
Visualizzazione Riepilogo del modello nel riquadro principale



La visualizzazione Riepilogo del modello mostra un'istantanea (o riepilogo) del modello di cluster, compresa una misura della silhouette di coesione e separazione dei cluster, che è ombreggiata per indicare risultati scarsi, discreti o buoni. Questa istantanea consente di verificare rapidamente se la qualità è scarsa, nel qual caso è possibile decidere di tornare al nodo per la creazione dei modelli per correggere le impostazioni del modello di cluster e ottenere un risultato migliore.

La qualità del risultato (scarso, discreto, buono) è basata sul lavoro di Kaufman e Rousseeuw (1990) relativo all'interpretazione delle strutture dei cluster. Nella visualizzazione Riepilogo del modello, un risultato buono equivale a quei dati che rispecchiano la classificazione di Kaufman e Rousseeuw di ragionevole o forte indizio di una struttura di cluster, un risultato discreto rispecchia la classificazione di indizio debole, un risultato scarso corrisponde alla classificazione di assenza di indizio significativo.

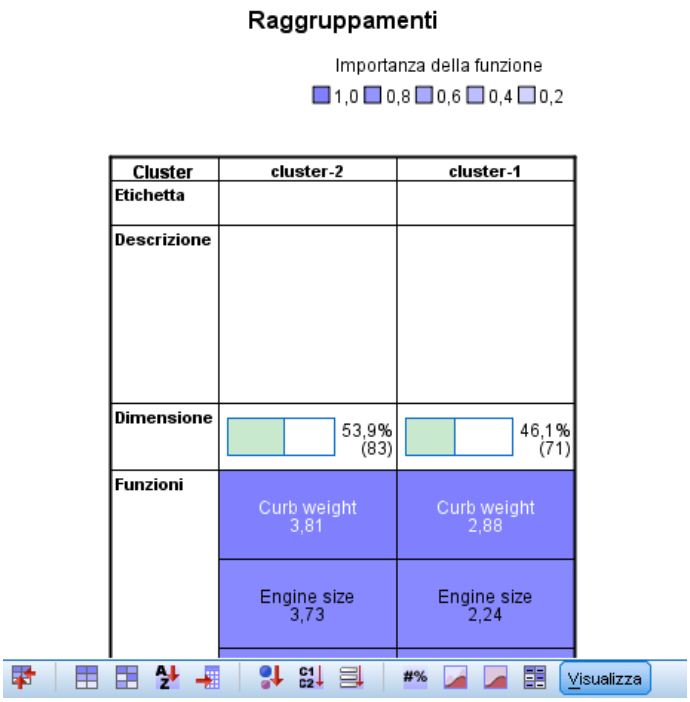
La misura della silhouette viene calcolata su tutti i record, (B−A) / max(A,B), dove A è la distanza del record dal centro del relativo cluster e B è la distanza del record dal centro del più vicino cluster a cui non appartiene. Un coefficiente di silhouette pari a 1 indica che tutti i casi si trovano direttamente in corrispondenza dei centri dei relativi cluster. Un valore pari a −1 indica che tutti i casi si trovano in corrispondenza dei centri di altri cluster. Il valore 0 indica, in media, che i casi sono equidistanti tra il centro del rispettivo cluster e il cluster più vicino.

Il riepilogo include una tabella che contiene le informazioni seguenti:

- **Algoritmo.** L'algoritmo di raggruppamento utilizzato (ad esempio, "TwoStep").
- **Funzioni di input.** Il numero di campi, noti anche come **input** o **predittori**.
- **Raggruppamenti.** Il numero di cluster nella soluzione.

Visualizzazione cluster

Figura 24-6
Visualizzazione Centri cluster nel riquadro principale



La visualizzazione Cluster contiene una griglia cluster-per-funzioni che comprende il nome, le dimensioni e il profilo di ciascun cluster.

Le colonne della griglia contengono le seguenti informazioni:

- **Cluster.** I numeri di cluster creati dall'algoritmo.
- **Etichetta.** L'eventuale etichetta applicata a ciascun cluster (che è vuota, per impostazione predefinita). Fare doppio clic nella cella per immettere un'etichetta che descrive il contenuto del cluster: ad esempio, "Acquirenti di auto di lusso".
- **Descrizione.** L'eventuale descrizione del contenuto del cluster (che è vuota, per impostazione predefinita). Fare doppio clic nella cella per immettere una descrizione del cluster: ad esempio, "età oltre i 55 anni, professionisti, reddito superiore a 100.000 euro".
- **Dimensioni.** Le dimensioni di ciascun cluster sotto forma di percentuale dell'intero campione di cluster. Ogni cella relativa alle dimensioni all'interno della griglia visualizza una barra verticale che mostra la percentuale delle dimensioni all'interno del cluster, la percentuale delle dimensioni in formato numerico e il conteggio dei casi di cluster.
- **Funzioni.** I singoli input o predittori, ordinati per impostazione predefinita in base all'importanza globale. Se delle colonne hanno dimensioni uguali vengono mostrate in base ai numeri di cluster in ordine crescente.

L'importanza generale di una funzione è indicata dal colore dell'ombreggiatura di sfondo della cella; la funzione più importante è la più scura, mentre quella meno importante è priva di ombreggiatura. Una guida al di sopra della tabella indica l'importanza associata al colore di ciascuna cella relativa a una funzione.

Quando si passa il mouse sopra una cella, vengono visualizzati il nome completo o l'etichetta della funzione e il valore di importanza della cella. È possibile che vengano visualizzate altre informazioni, a seconda della visualizzazione e del tipo di funzione. Nella visualizzazione Centri cluster, si tratta della statistica della cella e del valore della cella; ad esempio: "Media: 4.32". Per le funzioni categoriali la cella mostra il nome della categoria (modale) più frequente e la relativa percentuale.

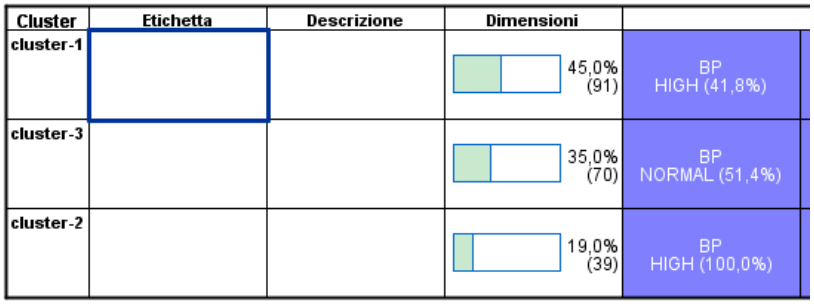
All'interno della visualizzazione dei cluster, è possibile selezionare diversi metodi per visualizzare le informazioni sul cluster:

- Trasponi cluster e funzioni. Per ulteriori informazioni, vedere l'argomento Trasponi cluster e funzioni a pag. 177.
- Ordina funzioni. Per ulteriori informazioni, vedere l'argomento Ordina funzioni a pag. 178.
- Ordina cluster. Per ulteriori informazioni, vedere l'argomento Ordina cluster a pag. 178.
- Seleziona contenuto celle. Per ulteriori informazioni, vedere l'argomento Contenuti cella a pag. 178.

Trasponi cluster e funzioni

Per impostazione predefinita, i cluster sono visualizzati come colonne e le funzioni come righe. Per invertire questa modalità, fare clic sul pulsante Trasponi cluster e funzioni a sinistra dei pulsanti Ordina funzioni in base a. Ad esempio, è possibile utilizzare questa opzione quando sono visualizzati troppi cluster, per ridurre la quantità di scorrimento orizzontale necessario per visionare i dati.

Figura 24-7
Cluster trasposti nel riquadro principale



Ordina funzioni

I pulsanti Ordina funzioni in base a consentono di selezionare il modo in cui sono visualizzate le celle delle funzioni:

- **Importanza globale.** È l'impostazione predefinita. Le funzioni vengono organizzate in ordine di importanza globale decrescente, e l'ordinamento è lo stesso tra i cluster. Se in qualche funzione sono presenti dei valori di importanza a pari merito, le funzioni a pari merito vengono elencate in ordine crescente in base ai nomi delle funzioni stesse.
- **Importanza entro i cluster.** Le funzioni vengono ordinate rispetto alla loro importanza per ciascun cluster. Se in qualche funzione sono presenti dei valori di importanza a pari merito, le funzioni a pari merito vengono elencate in ordine crescente in base ai nomi delle funzioni stesse. Quando si seleziona questa opzione, di solito l'ordine varia tra i cluster.
- **Nome.** Le funzioni vengono ordinate alfabeticamente in base al nome.
- **Ordine dei dati.** Le funzioni vengono ordinate in base al loro ordine nell'insieme di dati.

Ordina cluster

Per impostazione predefinita, i cluster vengono ordinati in modo decrescente in base alla dimensione. I pulsanti Ordina cluster in base a consentono di ordinarli alfabeticamente per nome o, se sono state create delle etichette alfanumeriche univoche, rispetto a queste ultime.

Le funzioni con la stessa etichetta vengono ordinate in base al nome del cluster. Se i cluster sono ordinati in base alle etichette e si modifica l'etichetta di un cluster, l'ordinamento viene aggiornato automaticamente.

Contenuti cella

I pulsanti Celle consentono di modificare la visualizzazione dei contenuti delle celle per quanto riguarda le funzioni e i campi di valutazione.

- **Centri cluster.** Per impostazione predefinita, le celle visualizzano i nomi e le etichette delle funzioni e la tendenza centrale per ciascuna combinazione cluster/funzione. La media viene mostrata per i campi continui e la moda (categoria che ricorre più frequentemente) con la percentuale della categoria per i campi categoriali.
- **Distribuzioni assolute.** Mostra i nomi e le etichette e le distribuzioni assolute delle funzioni all'interno di ciascun cluster. Per le funzioni categoriali, la schermata visualizza dei grafici a barre a cui sono sovrapposte delle categorie ordinate in modo crescente rispetto ai valori dei dati. Per le funzioni continue, la schermata mostra un grafico di densità regolare che utilizza gli stessi punti finali e intervalli per ciascun cluster.

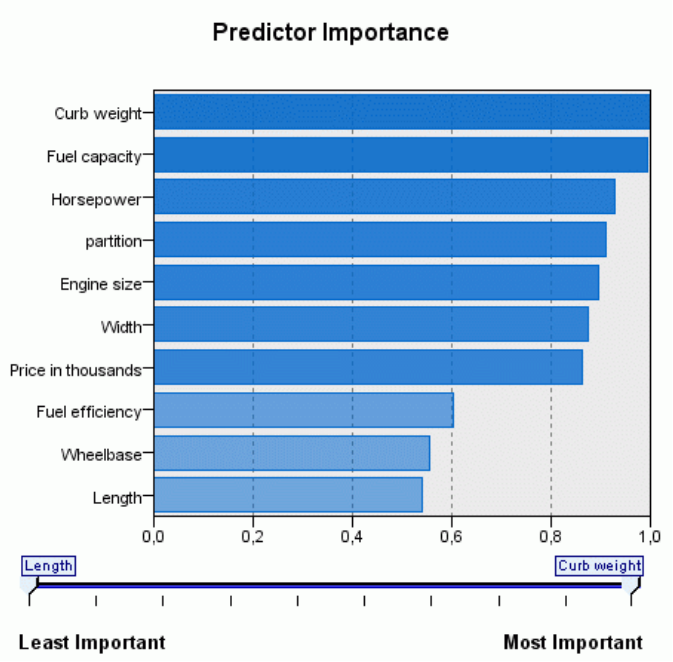
  La schermata in rosso pieno mostra la distribuzione dei cluster, mentre quella più chiara rappresenta i dati globali.
- **Distribuzioni relative.** Mostra i nomi e le etichette delle funzioni e le distribuzioni relative nelle celle. In generale, le schermate sono simili a quelle visualizzate per le distribuzioni assolute, a eccezione del fatto che vengono mostrate le distribuzioni relative.

La schermata in rosso pieno mostra la distribuzione dei cluster, mentre quella più chiara rappresenta i dati globali.

- **Visualizzazione di base.** In presenza di molti cluster, può risultare difficile visualizzare i dettagli senza ricorrere allo scorrimento. Per ridurre la quantità di scorrimento, selezionare questa visualizzazione per passare a una versione più compatta della tabella.

Visualizzazione Importanza predittore nei cluster

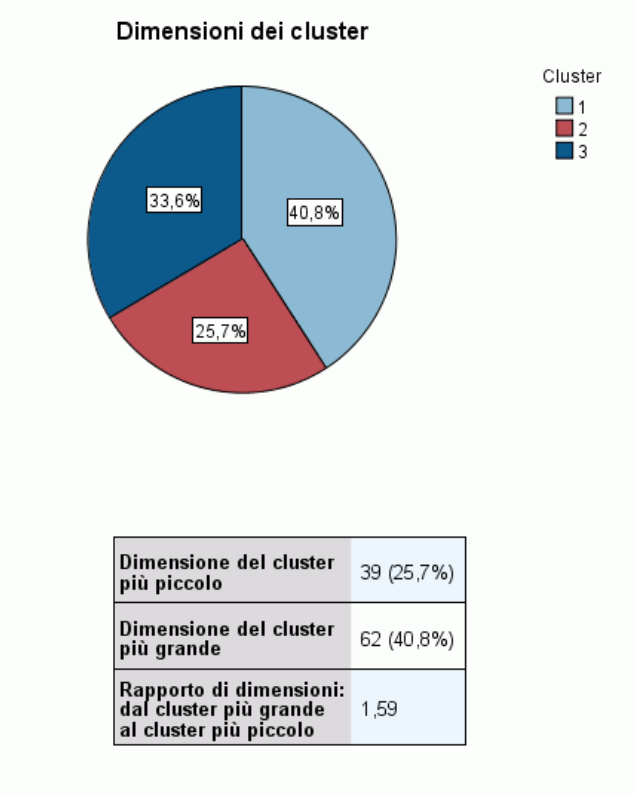
Figura 24-8
Visualizzazione Importanza predittore nei cluster nel riquadro collegato



La visualizzazione Importanza predittore mostra l'importanza relativa di ciascun campo nella stima del modello.

Visualizzazione Dimensioni dei cluster

Figura 24-9
La visualizzazione Dimensioni dei cluster nel riquadro collegato



Dimensione del cluster più piccolo	39 (25,7%)
Dimensione del cluster più grande	62 (40,8%)
Rapporto di dimensioni: dal cluster più grande al cluster più piccolo	1,59

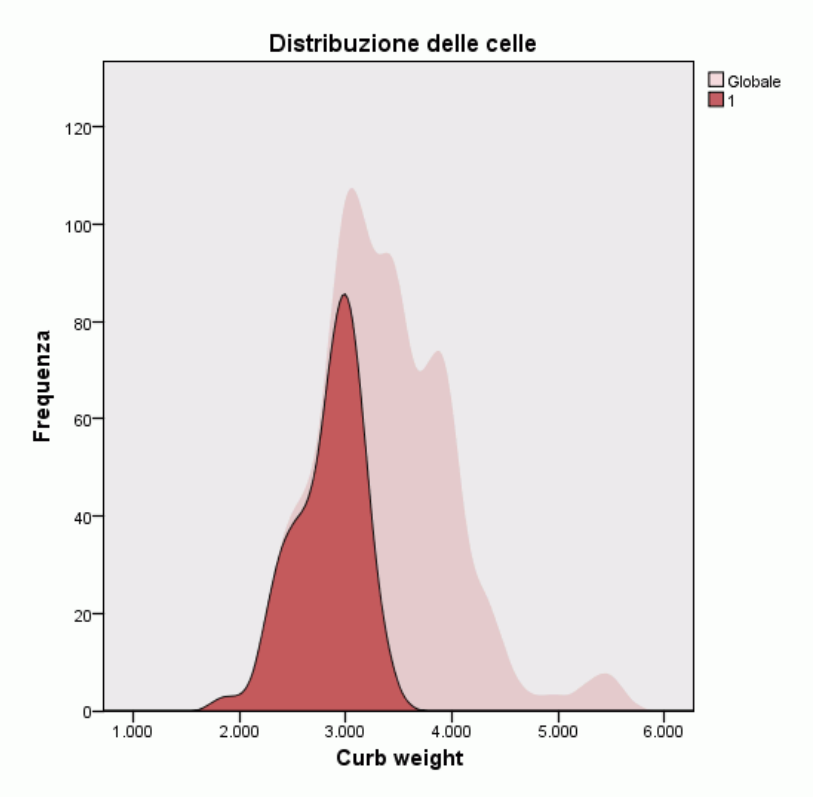
La visualizzazione Dimensioni dei cluster mostra un grafico a torta che contiene ciascun cluster. La dimensione percentuale di ciascun cluster viene mostrata in ogni fetta; passare il mouse sopra ogni fetta per visualizzare il conteggio al suo interno.

Al di sotto del grafico, una tabella elenca le seguenti informazioni relative alle dimensioni:

- La dimensione del cluster più piccolo (sia il conteggio che una percentuale rispetto al totale).
- La dimensione del cluster più grande (sia il conteggio che una percentuale rispetto al totale).
- Il rapporto tra la dimensione del cluster più grande e quella del cluster più piccolo.

Visualizzazione Distribuzione delle celle

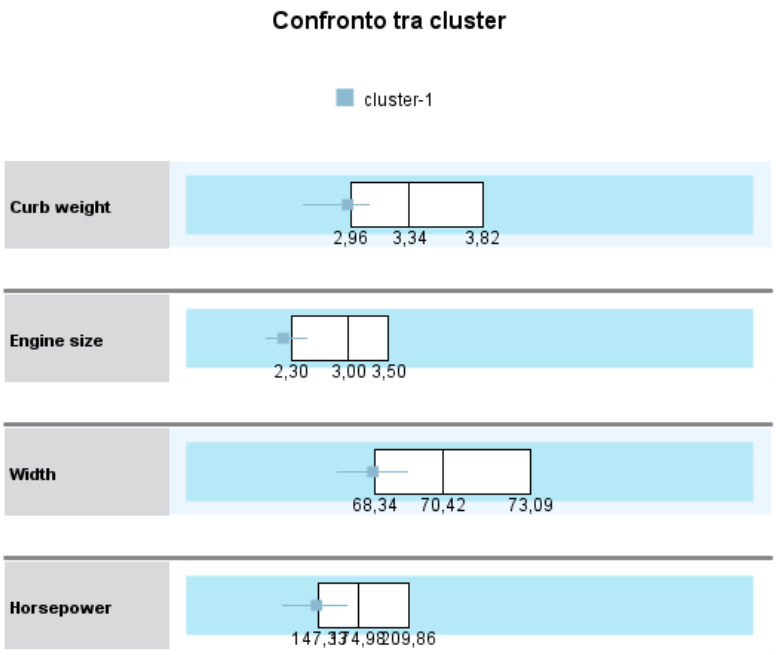
Figura 24-10
Visualizzazione Distribuzione delle celle nel riquadro collegato



La visualizzazione Distribuzione delle celle mostra un grafico espanso e più dettagliato della distribuzione dei dati per qualsiasi cella di funzione selezionata nella tabella del riquadro principale dei cluster.

Visualizzazione Confronto tra cluster

Figura 24-11
Visualizzazione Confronto tra cluster nel riquadro collegato



La visualizzazione Confronto tra cluster è costituita da un layout a griglia, con le funzioni nelle righe e i cluster selezionati nelle colonne. Questa visualizzazione aiuta a comprendere meglio i fattori che formano i cluster; inoltre, consente di visualizzare le differenze tra i cluster non solo confrontandoli con i dati globali ma anche l'uno con l'altro.

Per selezionare i cluster da visualizzare, fare clic sulla parte superiore della colonna dei cluster nel riquadro principale Cluster. Fare clic tenendo premuto Ctrl o Maiusc per selezionare o deselezionare più di un cluster per il confronto.

Nota: è possibile selezionare un massimo di cinque cluster per la visualizzazione.

I cluster vengono mostrati nell'ordine in cui sono stati selezionati, mentre l'ordine dei campi è determinato dall'opzione Ordina funzioni in base a. Quando si seleziona Importanza entro i cluster, i campi vengono sempre ordinati in base all'importanza globale.

I grafici sullo sfondo mostrano le distribuzioni globali di ciascuna funzione:

- Le funzioni categoriali vengono visualizzate sotto forma di grafici a punti, dove la dimensione del punto indica la categoria più frequente/modale per ogni cluster (per funzione).
- Le funzioni continue vengono visualizzate sotto forma di grafici a scatole, che mostrano le mediane globali e le distanze interquartiliche.

Sovrapposti a queste visualizzazioni in secondo piano sono i grafici a scatole per i cluster selezionati:

- Per le funzioni continue, i simboli a punta quadrata e le linee orizzontali indicano la mediana e la distanza interquartilica per ciascun cluster.
- Ciascun cluster è rappresentato per mezzo di un colore diverso, mostrato nella parte superiore della visualizzazione.

Esplorazione del Visualizzatore cluster

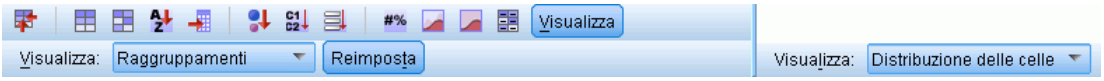
Il Visualizzatore cluster è una schermata interattiva, che consente di:

- Selezionare un campo o un cluster per visualizzare ulteriori dettagli.
- Confrontare i cluster per selezionare gli elementi desiderati.
- Modificare la visualizzazione.
- Trasporre gli assi.

Utilizzo delle barre degli strumenti

Le informazioni visualizzate nei riquadri destro e sinistro possono essere controllate mediante le opzioni delle barre degli strumenti. I controlli delle barre degli strumenti consentono di modificare l'orientamento della visualizzazione (dall'alto verso il basso, da sinistra verso destra o da destra verso sinistra). Inoltre, è anche possibile reimpostare le impostazioni predefinite del visualizzatore e aprire una finestra di dialogo per specificare il contenuto della visualizzazione Cluster nel riquadro principale.

Figura 24-12
Barre degli strumenti per il controllo dei dati visualizzati nel Visualizzatore cluster



Le opzioni Ordina funzioni in base a, Ordina cluster in base a, Celle e Visualizza sono disponibili solo quando si seleziona la visualizzazione Cluster nel riquadro principale. Per ulteriori informazioni, vedere l'argomento Visualizzazione cluster a pag. 176.

	Vedere Trasponi cluster e funzioni a pag. 177
	Vedere Ordina funzioni in base a a pag. 178
	Vedere Ordina cluster in base a a pag. 178
	Vedere Celle a pag. 178

Controllo della visualizzazione cluster

Per controllare ciò che viene mostrato nella visualizzazione Cluster nel riquadro principale, fare clic sul pulsante Visualizza; viene aperta la finestra di dialogo Visualizza.

Figura 24-13
Visualizzatore cluster - Opzioni di visualizzazione

Funzioni. Selezionata per impostazione predefinita. Per nascondere tutte le funzioni di input, deselezionare la casella di controllo.

Campi di valutazione. Scegliere i campi di valutazione (campi non utilizzati per creare il modello di cluster, ma inviati al Viewer modelli per valutare i cluster) da visualizzare; per impostazione predefinita non ne è visualizzato nessuno. *Nota*: questa casella di controllo non è disponibile se non è disponibile alcun campo di valutazione.

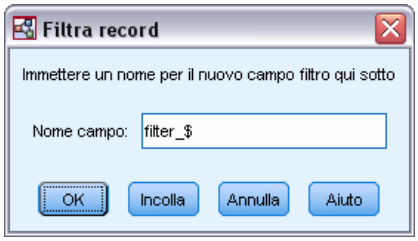
Descrizioni cluster. Selezionata per impostazione predefinita. Per nascondere tutte le celle delle descrizioni dei cluster, deselezionare la casella di controllo.

Dimensioni dei cluster. Selezionata per impostazione predefinita. Per nascondere tutte le celle delle dimensioni dei cluster, deselezionare la casella di controllo.

Numero massimo di categorie. Specifica il numero massimo di categorie da visualizzare nei grafici delle funzioni categoriali; il valore predefinito è 20.

Filtraggio dei record

Figura 24-14
Visualizzatore cluster - Filtra casi



Per maggiori informazioni sui casi in un particolare cluster o gruppo di cluster, selezionare un sottoinsieme di record per un'ulteriore analisi basata sui cluster selezionati.

- Selezionare i cluster nella visualizzazione Cluster del Visualizzatore cluster. Fare clic tenendo premuto il tasto Ctrl per selezionare più cluster.

- Dai menu, scegliere:
  Genera > Filtra record...

- Inserire un nome per la variabile di filtro. I record dei cluster selezionati riceveranno un valore pari a 1 per il campo. Tutti gli altri record riceveranno un valore pari a 0 e saranno esclusi dalle analisi successive fino a quando non viene modificato lo stato del filtro.

- Fare clic su OK.

Capitolo
25

Cluster gerarchico

Questa procedura consente di identificare gruppi di casi relativamente omogenei in base alle caratteristiche selezionate, utilizzando un algoritmo che inizia con ciascun caso (o variabile) in un cluster distinto e che combina i cluster fino a quando ne rimane solo uno. È possibile analizzare le variabili semplici oppure scegliere una delle trasformazioni di standardizzazione disponibili. Le misure di similarità e dissimilarità vengono generate dalla procedura Distanze. A ciascun livello verranno visualizzate statistiche in base alle quali selezionare la soluzione migliore.

Esempio. Esistono gruppi di trasmissioni televisive identificabili che attraggono tipi di audience analoghi all'interno di ciascun gruppo? Utilizzando la cluster gerarchica è possibile raggruppare le trasmissioni televisive (casi) in gruppi omogenei in base alle caratteristiche degli spettatori. Questo metodo può essere utilizzato per identificare i segmenti di mercato. In alternativa, è possibile raggruppare le città (casi) in gruppi omogenei in modo che da poter selezionare città con caratteristiche confrontabili per verificare diverse strategie di mercato.

Statistiche. Programma di agglomerazione, matrice delle distanze (o similarità) e cluster di appartenenza per un'unica soluzione o una serie di soluzioni. Grafici: dendrogrammi e grafici a stalattite

Dati. Le variabili possono essere quantitative, binarie o dati di conteggio. Lo scaling delle variabili è molto importante in quanto le differenze di scaling possono influire sulle soluzioni dei cluster. Se lo scaling delle variabili presenta differenze notevoli (ad esempio, una variabile viene misurata in dollari e l'altra in anni), è consigliabile standardizzarle. Ciò può essere effettuato in modo automatico mediante la procedura Cluster gerarchica.

Ordine dei casi. Se le distanze assegnate o le similarità sono presenti nei dati iniziali o nei cluster aggiornati durante l'unione, la soluzione del cluster risultante può essere influenzata dall'ordine dei casi del file. Può essere utile ottenere più soluzioni diverse con casi disposti in ordini casuali diversi per verificare la stabilità di una soluzione specifica.

Assunzioni. Le misure di dissimilarità o di similarità utilizzate devono essere idonee per i dati analizzati. Per ulteriori informazioni sulla scelta delle misure di dissimilarità e similarità, vedere la procedura Distanze. È inoltre necessario includere nell'analisi tutte le variabili significative. L'omissione di variabili importanti può portare a soluzioni improprie. Poiché la cluster gerarchica rappresenta un metodo esplorativo, i risultati devono essere considerati provvisori finché non vengano confermati da un campione indipendente.

Per ottenere una cluster gerarchica

- Dai menu, scegliere:
  Analizza > Classifica > Cluster gerarchico...

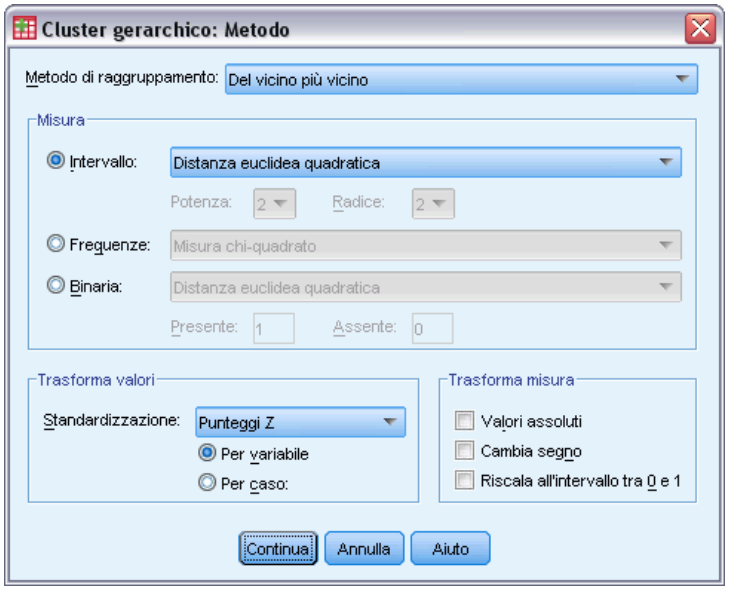
Figura 25-1
Finestra di dialogo Cluster gerarchica

- Per raggruppare i casi in cluster è necessario selezionare almeno una variabile numerica. Per raggruppare le variabili in cluster è necessario selezionare almeno tre variabili numeriche.

  È inoltre possibile selezionare una variabile di identificazione per etichettare i casi.

Cluster gerarchica: Metodo

Figura 25-2
Finestra di dialogo Cluster gerarchica: Metodo



Metodo di raggruppamento. Le alternative disponibili sono: Legame medio fra i gruppi, Legame medio entro gruppi, Del vicino più vicino, Del vicino più lontano, Centroide, Mediana e Ward.

Misura. Consente di specificare la misura di similarità o dissimilarità da utilizzare per il raggruppamento. Selezionare il tipo di dati e la misura di similarità o dissimilarità desiderata:

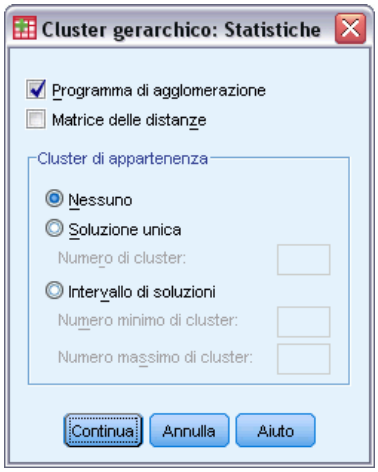
- **Intervallo.** Le alternative disponibili sono: Distanza euclidea, Distanza euclidea quadratica, Coseno, Correlazione di Pearson, Chebychev, City-Block, Minkowski e Personalizzato.
- **Conteggi.** Le alternative disponibili sono: Misura chi-quadrato e Misura phi-quadrato.
- **Binaria.** Le alternative disponibili sono: Distanza euclidea, Distanza euclidea quadratica, Differenza di dimensione, Differenza di modello, Varianza, Dispersione, Forma, Corrispondenza semplice, Correlazione phi a 4 punti, Lambda, *D* di Anderberg, Dice, Hamann, Jaccard, Kulczynski 1, Kulczynski 2, Lance e Williams, Ochiai, Rogers e Tanimoto, Russel e Rao, Sokal e Sneath 1, Sokal e Sneath 2, Sokal e Sneath 3, Sokal e Sneath 4, Sokal e Sneath 5, *Y* di Yule e *Q*di Yule.

Trasforma valori. Consente di standardizzare i valori dei dati per casi o valori prima di calcolare le similarità (non disponibile per i dati binari). I metodi di standardizzazione disponibili sono: punteggi *z*, intervallo da −1 a 1, intervallo da 0 a 1, ampiezza massima di 1, media di 1 e deviazione standard di 1.

Trasforma misure. Consente di trasformare i valori generati dalla misura di distanza. Questi verranno applicati dopo il calcolo della misura di distanza. Le alternative disponibili sono: Valori assoluti, Cambia segno e Riscala all'intervallo 0–1.

Cluster gerarchica: Statistiche

Figura 25-3
Finestra di dialogo Cluster gerarchica: Statistiche



Programma di agglomerazione. Consente di visualizzare i casi o i cluster accorpati ad ogni stadio, le distanze tra i casi o i cluster da accorpare e l'ultimo livello di cluster in cui un caso (o una variabile) è stato accorpato al cluster.

Matrice delle distanze. Fornisce le distanze o le similarità tra gli elementi.

Cluster di appartenenza. Viene visualizzato il cluster a cui viene assegnato ciascun caso a uno o più stadi della combinazione dei cluster. Le opzioni disponibili sono Soluzione unica e Intervallo di soluzioni.

Cluster gerarchica: Grafici

Figura 25-4
Finestra di dialogo Cluster gerarchica: Grafici

Dendrogramma. Viene visualizzato un **dendrogramma**. Utilizzando i dendrogrammi è possibile valutare la coesione dei cluster formati ed ottenere informazioni sul numero di cluster che è opportuno tenere.

A stalattite. Visualizza un **grafico a stalattite**, inclusi tutti i cluster o l'intervallo di cluster specificato. Nei grafici a stalattite vengono visualizzate informazioni sulle modalità con cui i casi vengono combinati in cluster ad ogni iterazione dell'analisi. Specificando l'orientamento desiderato è possibile selezionare un grafico verticale o orizzontale.

Cluster gerarchica: Salva nuove variabili

Figura 25-5
Finestra di dialogo Cluster gerarchica: Salva

Cluster di appartenenza. Consente di salvare i cluster di appartenenza per una soluzione unica o per un intervallo di soluzioni. Le variabili salvate possono essere utilizzate in analisi successive per valutare altre differenze tra i gruppi.

Funzioni aggiuntive della sintassi del comando CLUSTER

La procedura Cluster gerarchica usa la sintassi del comando `CLUSTER`. Il linguaggio della sintassi dei comandi consente inoltre di:

- Usare più metodi di raggruppamento in una singola analisi.
- Leggere ed analizzare una matrice di prossimità.
- Scrivere una matrice di prossimità sul disco per analizzarla in seguito.
- Specificare i valori per la potenza e la radice nella misura della distanza personalizzata (potenza).
- Specificare i nomi delle variabili salvate.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
26
Cluster con metodo delle k-medie

Questa procedura consente di identificare gruppi di casi relativamente omogenei in base alle caratteristiche selezionate, utilizzando un algoritmo in grado di gestire un elevato numero di casi. Tale algoritmo, tuttavia, richiede l'indicazione del numero di cluster. È possibile specificare i centri iniziali del cluster, se si conosce questa informazione. È possibile selezionare uno dei due metodi disponibili per la classificazione dei casi, ovvero l'aggiornamento iterativo dei centri cluster oppure la semplice classificazione. È possibile salvare l'appartenenza al cluster, le informazioni sulla distanza e i centri del cluster finali. È inoltre possibile specificare una variabile i cui valori possono essere utilizzati per etichettare l'output caso per caso. i può inoltre richiedere l'analisi delle statistiche *F* di varianza. Se da un lato queste statistiche sono opportunistiche, ovvero vengono eseguiti tentativi di raggruppamenti che presentino differenze, le corrispondenti dimensioni relative forniscono informazioni sul contributo apportato da ciascuna variabile alla separazione dei gruppi.

Esempio. Quali sono i gruppi di show televisivi che attraggono un pubblico analogo all'interno di ciascun gruppo? Il metodo cluster *k*-medie consente di raggruppare gli show televisivi (casi) in*k* gruppi omogenei in base alle caratteristiche degli spettatori. Questo processo può essere utilizzato per identificare i segmenti di mercato. In alternativa, è possibile raggruppare le città (casi) in gruppi omogenei in modo che da poter selezionare città con caratteristiche confrontabili per verificare diverse strategie di mercato.

Statistiche. Soluzione completa: centri iniziali del cluster, tabella ANOVA. Per ciascun caso: informazioni sui cluster, distanza dal centro del cluster.

Dati. Le variabili devono essere quantitative a livello di intervallo o di rapporto. Se le variabili sono binarie o conteggi, utilizzare la procedura Cluster gerarchica.

Ordine dei casi e dei centri di cluster iniziale. L'algoritmo predefinito per la scelta dei centri di cluster iniziali varia a seconda dell'ordine dei casi. L'opzioneUsa medie mobili della finestra di dialogo Iterazioni rende la soluzione risultante potenzialmente dipendente dall'ordine dei casi, indipendentemente dai centri di cluster scelti inizialmente. Se si utilizza uno di questi metodi, può essere utile ottenere più soluzioni diverse con casi disposti in ordini casuali diversi per verificare la stabilità di una soluzione specifica. Per evitare problemi con l'ordine dei casi, è consigliabile specificare i centri di cluster iniziali ed evitare di usare l'opzione Usa medie mobili. Tuttavia, l'ordinamento dei centri di cluster iniziali può influire sulla soluzione se esistono distanze assegnate dai casi ai centri di cluster. Per valutare la stabilità di una soluzione, è possibile confrontare i risultati delle analisi con diverse permutazioni dei valori dei centri iniziali.

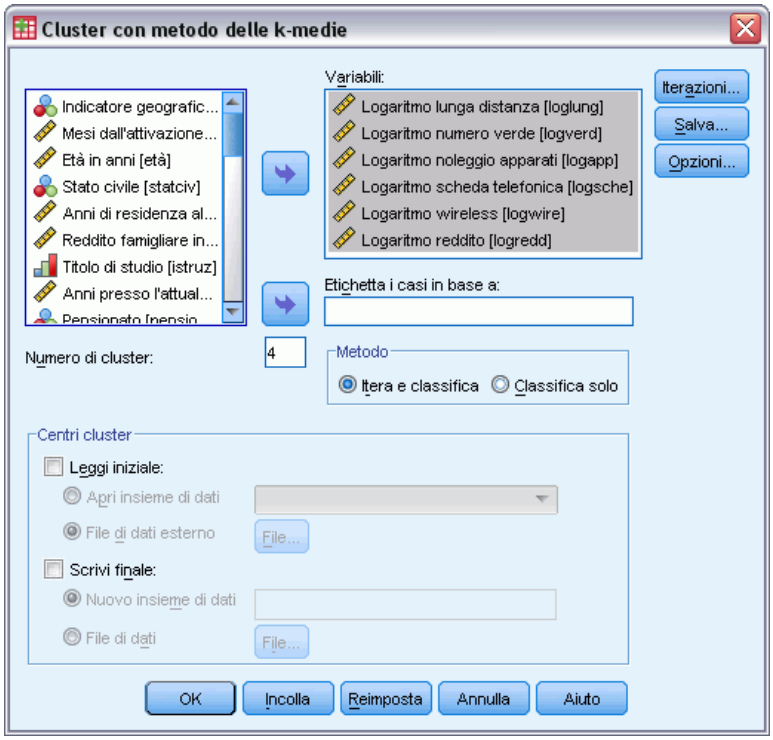
Assunzioni. Le distanze vengono calcolate utilizzando la distanza euclidea semplice. Se si desidera utilizzare un'altra misura di distanza o di similarità, utilizzare la procedura Cluster gerarchica. La scalatura delle variabili è un'operazione che deve essere effettuata con molta attenzione. Se le variabili vengono misurate con scale diverse (ad esempio se una variabile è espressa in dollari e un'altra è espressa in anni), i risultati possono essere fuorvianti. In questi casi è consigliabile standardizzare le variabili prima di procedere con l'analisi cluster *k*medie (utilizzando la procedura Descrittive). Questa procedura presume che sia stato selezionato il numero esatto di cluster e che

siano state incluse tutte le variabili rilevanti. Se è stato selezionato un numero di cluster inesatto o sono state omesse variabili importanti, i risultati possono essere inattendibili.

Per ottenere un'analisi cluster K-medie

- Dai menu, scegliere:
  Analizza > Classifica > Cluster k-medie...

Figura 26-1
Finestra di dialogo Cluster K-medie



- Selezionare le variabili da utilizzare nell'analisi cluster.
- Specificare il numero di cluster. Il numero di cluster specificato deve essere almeno di 2 e non deve essere maggiore al numero di casi del file dati.
- Selezionare il metodo Itera e classifica oppure il metodo Classifica soltanto.
- In alternativa, selezionare una variabile di identificazione per etichettare i casi.

Efficienza dell'analisi cluster K-medie

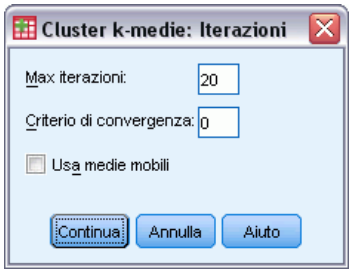
Il comando Cluster *k*-medie è efficace principalmente in quanto non calcola le distanze tra tutte le coppie di casi, a differenza di numerosi algoritmi di raggruppamento, ad esempio quello utilizzato dal comando per la Cluster gerarchica di SPSS.

Per ottenere la massima efficienza, creare un campione di casi e utilizzare il metodo Itera e classifica per determinare i centri cluster. Selezionare Scrivi valori finali su file. Quindi, ripristinare tutto il file di dati e selezionare Classifica soltanto come metodo e selezionare Leggi valori iniziali per

classificare tutto il file utilizzando i centri valutati per il campione. È possibile leggere o scrivere da un file o file di dati. I file di dati possono anche essere riutilizzati nella stessa sessione, ma non vengono salvati come file a meno che siano stati salvati come tali alla fine della sessione. I nomi degli insiemi di dati devono essere conformi alle regole di denominazione delle variabili.

Cluster K-medie: Iterazioni

Figura 26-2
Finestra di dialogo Cluster K-medie: Iterazioni



Nota: queste opzioni sono disponibili solo se si seleziona il metodo Itera e classifica nella finestra di dialogo Cluster con metodo delle K-medie.

Massimo numero di iterazioni. Consente di impostare il numero massimo di iterazioni per l'algoritmo *k*-medie. Le iterazioni si interromperanno al numero impostato, anche se il criterio di convergenza non viene soddisfatto. Il numero deve essere compreso tra 1 e 999.

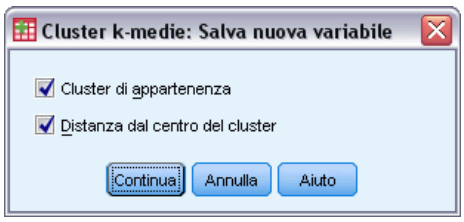
Per riprodurre l'algoritmo utilizzato dal comando Quick Cluster delle versioni di SPSS precedenti alla 5.0, impostare l'opzione Massimo numero di iterazioni su 1.

Criterio di convergenza. Determina il termine dell'iterazione. Rappresenta una proporzione della distanza minima fra i centri iniziali del cluster in modo che sia maggiore di 0 e minore di 1. Se, ad esempio, il criterio è 0,02, il processo di iterazione terminerà quando un'iterazione completa non è in grado di spostare i centri cluster di una distanza maggiore del 2% della distanza minima fra i centri iniziali del cluster.

Usa medie mobili. Consente di richiedere l'aggiornamento dei centri cluster in seguito all'assegnazione di ciascun caso Se non viene selezionata questa opzione, i nuovi centri del cluster verranno calcolati quando tutti i casi saranno stati assegnati.

Cluster K-medie: Salva

Figura 26-3
Finestra di dialogo Cluster K-medie: Salva nuove variabili



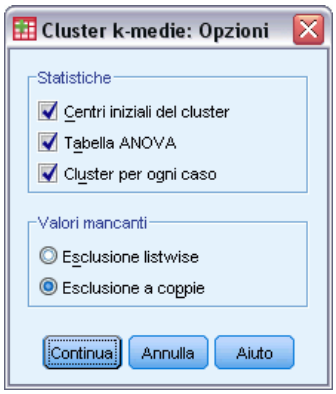
È possibile salvare informazioni sulla soluzione come nuove variabili da utilizzare in analisi successive:

Cluster di appartenenza. Consente di creare una nuova variabile che indica l'appartenenza finale al cluster di ciascun caso. I valori della nuova variabile sono compresi tra 1 e il numero di cluster.

Distanza dal centro. Consente di creare una nuova variabile che indica la distanza euclidea tra ciascun caso e il relativo centro di classificazione.

Cluster K-medie: Opzioni

Figura 26-4
Finestra di dialogo Cluster K-medie: Opzioni



Statistiche. È possibile selezionare le seguenti statistiche: centri iniziali del cluster, tabella ANOVA e informazioni sui cluster per ciascun caso.

- **Centri iniziali del cluster.** Prima stima delle medie delle variabili per ciascun cluster. In mancanza di indicazioni particolari, viene selezionato dai dati un numero di casi ben distanziati uguale al numero dei cluster. I centri dei cluster iniziali vengono usati per un primo ciclo di classificazione e poi vengono aggiornati.
- **Tabella ANOVA.** Visualizza una tabella di analisi della varianza con test F per ogni variabile di raggruppamento. I test F sono descrittivi e il livello di significatività fornisce informazioni utili. La tabella non viene creata se tutti i casi vengono assegnati a un solo cluster.
- **Cluster per ogni caso.** Visualizza per ogni caso il cluster di appartenenza finale e la distanza euclidea dal centro del cluster utilizzato per classificare il caso. Visualizza inoltre la distanza euclidea fra i centri finali.

Valori mancanti. Le opzioni disponibili sono Escludi casi listwise o Escludi casi pairwise.

- **Esclusione listwise.** Consente di escludere i casi coni valori mancanti per le variabili di raggruppamento dall'analisi.
- **Esclusione pairwise.** Consente di assegnare i casi ai cluster in base alle distanze calcolate da tutte le variabili con valori non mancanti.

Opzioni aggiuntive del comando QUICK CLUSTER

La procedura Cluster K-medie usa la sintassi del comando `QUICK CLUSTER`. Il linguaggio della sintassi dei comandi consente inoltre di:

- Accettare i primi *k* casi come centri dei cluster iniziali per evitare di dover leggere i dati normalmente usati per stimarli.
- Specificare i centri iniziali dei cluster direttamente come parte della sintassi del comando.
- Specificare i nomi delle variabili salvate.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
27

Test non parametrici

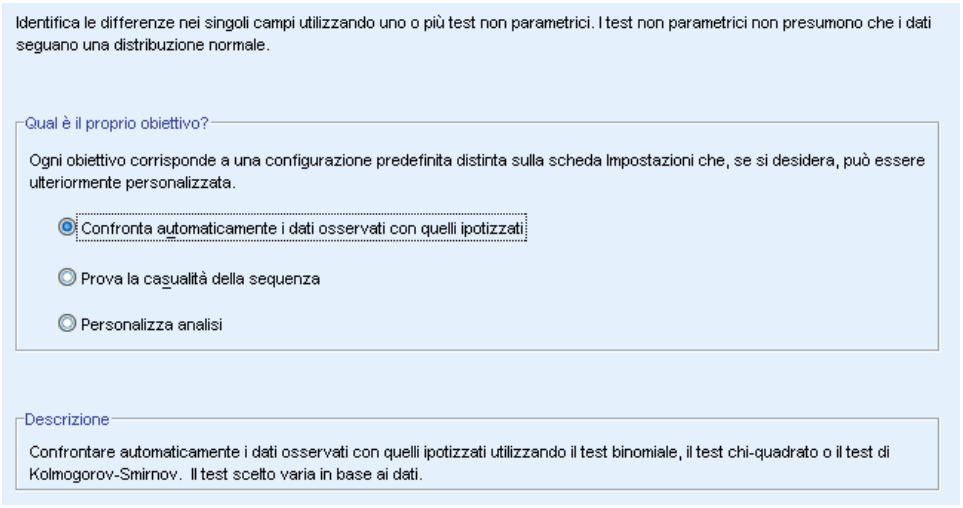
I test non parametrici formulano supposizioni minime sulla distribuzione sottostante dei dati. I test disponibili in queste finestre di dialogo si possono raggruppare in tre categorie generali a seconda di come sono organizzati i dati:

- Un test a campione singolo analizza un solo campo.
- Un test a campioni correlati confronta due o più campi per lo stesso insieme di casi.
- Un test a campioni indipendenti analizza un campo raggruppato secondo le categorie di un altro campo.

Test non parametrici a campione singolo

I test non parametrici a campione singolo identificano le differenze nei singoli campi utilizzando uno o più test non parametrici. I test non parametrici non presumono che i dati seguano una distribuzione normale.

Figura 27-1
Test non parametrici a campione singolo: scheda Obiettivo



Qual è il proprio obiettivo? Gli obiettivi consentono di specificare rapidamente varie impostazioni di uso comune per i test.

- **Confronta automaticamente i dati osservati con quelli ipotizzati.** Questo obiettivo applica il test binomiale ai campi categoriali con due sole categorie, il test chi-quadrato a tutti gli altri campi categoriali e il test di Kolmogorov-Smirnov ai campi continui.

- **Prova la casualità della sequenza.** Questo obiettivo utilizza il test delle successioni per verificare la casualità della sequenza di valori dei dati osservata.
- **Analisi personalizzata.** Selezionare questa opzione se si desidera modificare manualmente le impostazioni del test nella scheda Impostazioni. Si noti che questa impostazione viene selezionata automaticamente se in seguito si apportano modifiche incompatibili con l'obiettivo selezionato alle opzioni della scheda Impostazioni.

Per ottenere test non parametrici a campione singolo

Dai menu, scegliere:
Analizza > Test non parametrici > Campione singolo...

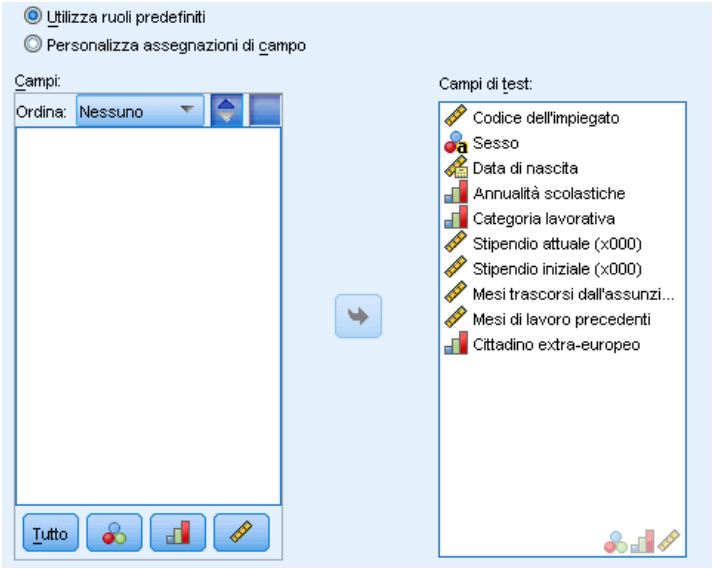
▶ Fare clic su Esegui.

Se lo si desidera, è possibile:

- Specificare un obiettivo nella scheda Obiettivi.
- Specificare le assegnazioni di campo nella scheda Campi.
- Specificare delle impostazioni avanzate nella scheda Impostazioni.

Scheda Campi

Figura 27-2
Test non parametrici a campione singolo: scheda Campi



La scheda Campi indica i campi che è necessario testare.

Utilizza ruoli predefiniti. Questa opzione utilizza le informazioni contenute nei campi esistenti. Tutti i campi con un ruolo predefinito come Input, Obiettivo o Entrambi saranno utilizzati come campi di test. È obbligatorio avere almeno un campo di test.

Utilizza assegnazioni campi personalizzate. Questa opzione consente di ignorare i ruoli dei campi. Dopo averla selezionata, compilare i campi riportati sotto:

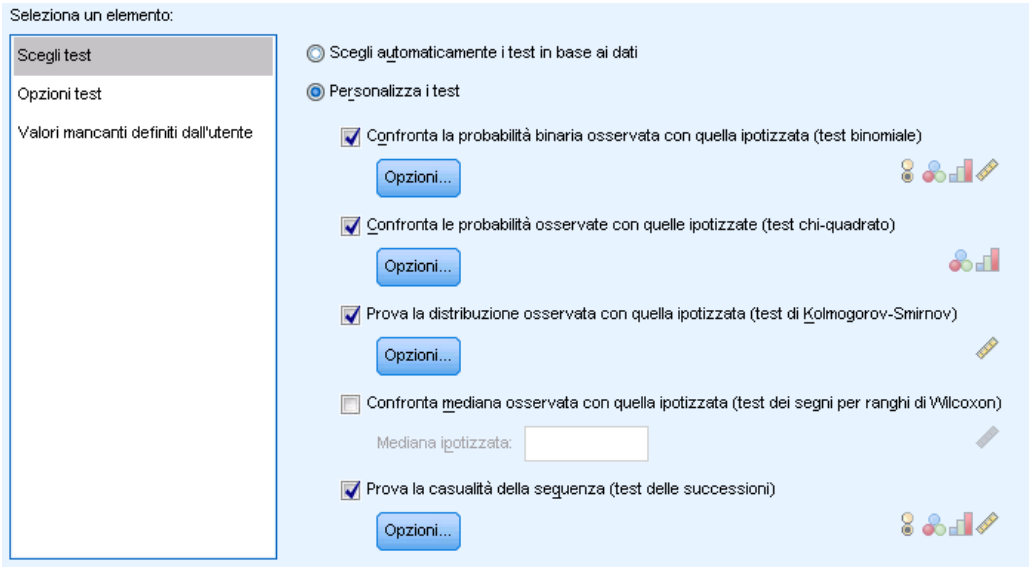
- **Campi di test.** Selezionare uno o più campi.

Scheda Impostazioni

La scheda Impostazioni contiene vari gruppi di impostazioni che è possibile modificare per perfezionare l'elaborazione dei dati da parte dell'algoritmo. Se si apportano modifiche alle impostazioni predefinite che risultano incompatibili con l'obiettivo selezionato, la scheda Obiettivo viene aggiornata automaticamente per selezionare l'opzione Personalizza analisi.

Scegli test

Figura 27-3
Test non parametrici a campione singolo: impostazioni di Scegli test



Queste impostazioni specificano i test da eseguire sui campi indicati nella scheda Campi.

Scegli automaticamente i test. Questa impostazione applica il test binomiale ai campi categoriali con due sole categorie valide (non mancanti), il test chi-quadrato a tutti gli altri campi categoriali e il test di Kolmogorov-Smirnov ai campi continui.

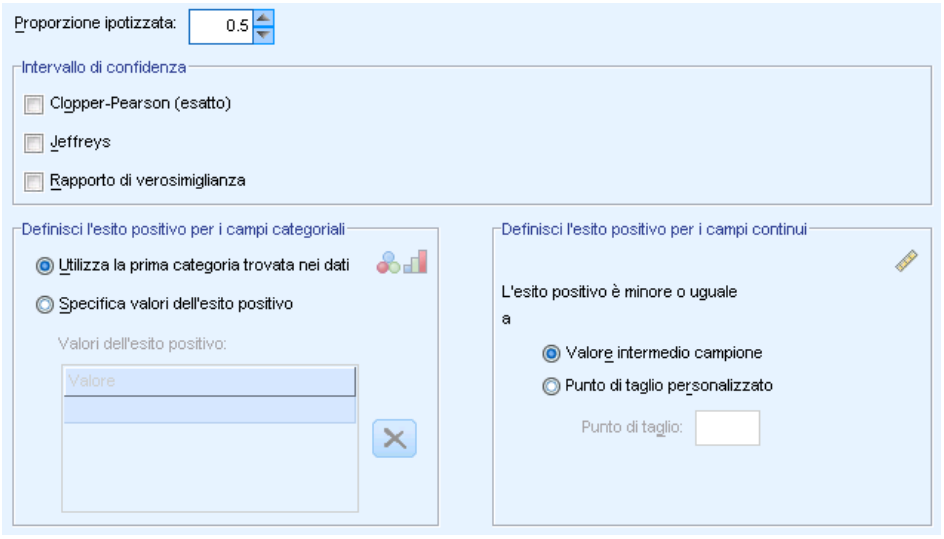
Personalizza i test. Questa impostazione consente di definire l'esecuzione di test specifici.

- **Confronta la probabilità binaria osservata con quella ipotizzata (test binomiale).** Il test binomiale può essere applicato a tutti i campi. Esso genera un test a campione singolo che verifica se la distribuzione osservata di un campo flag (campo categoriale con due sole categorie) è uguale a quella prevista da una distribuzione binomiale specificata. È possibile inoltre richiedere gli intervalli di confidenza. Vedere Test binomiale: Opzioni per informazioni più dettagliate sulle impostazioni dei test.

- **Confronta le probabilità osservate con quelle ipotizzate (test chi-quadrato).** Il test chi-quadrato viene applicato ai campi nominali e ordinali. Questa opzione genera un test a campione singolo che calcola una statistica chi-quadrato in base alle differenze tra le frequenze osservate e previste delle categorie di un campo. Vedere Test chi-quadrato: Opzioni per informazioni più dettagliate sulle impostazioni dei test.
- **Prova la distribuzione osservata con quella ipotizzata (test di Kolmogorov-Smirnov).** Il test di Kolmogorov-Smirnov viene applicato ai campi continui. Questa opzione genera un test a campione singolo che verifica se la funzione di distribuzione cumulata del campione di un campo è omogenea con una distribuzione uniforme, normale, di Poisson o esponenziale. Vedere Opzioni test di Kolmogorov-Smirnov per informazioni più dettagliate sulle impostazioni dei test.
- **Confronta mediana osservata con quella ipotizzata (test dei segni per ranghi di Wilcoxon).** Il test dei segni per ranghi di Wilcoxon viene applicato ai campi continui. Questa opzione genera un test a campione singolo del valore della mediana di un campo. Specificare un numero come mediana ipotizzata.
- **Prova la casualità della sequenza (test delle successioni).** Il test delle successioni viene applicato a tutti i campi. Questa opzione genera un test a campione singolo che verifica se la sequenza dei valori di un campo dicotomizzato è casuale. Vedere Test delle successioni: Opzioni per informazioni più dettagliate sulle impostazioni dei test.

Test binomiale: Opzioni

Figura 27-4
Test non parametrici a campione singolo: Test binomiale: Opzioni



Il test binomiale è destinato ai campi flag (campi categoriali con due sole categorie), ma viene applicato a tutti i campi utilizzando le regole per la definizione dell'"esito positivo".

Proporzione ipotizzata. Specifica la proporzione prevista di record definiti come "esito positivo" o *p*. Specificare un valore maggiore di 0 e minore di 1. Il valore predefinito è 0,5.

Intervallo di confidenza. Per calcolare gli intervalli di confidenza per i dati binari sono disponibili i seguenti metodi:

- **Clopper-Pearson (esatto).** Un intervallo esatto basato sulla distribuzione binomiale cumulata.
- **Jeffreys.** Un intervallo bayesiano basato sulla distribuzione a posteriori di *p* utilizzando la distribuzione a priori di Jeffreys.
- **Rapporto di verosimiglianza.** Un intervallo basato sulla funzione di verosimiglianza per *p*.

Definisci l'esito positivo per i campi categoriali. Specifica come viene definito l'"esito positivo" (il valore o i valori dei dati confrontati con la proporzione ipotizzata) per i campi categoriali.

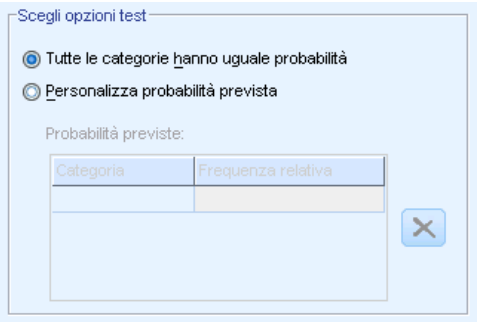
- Utilizza la prima categoria trovata nei dati esegue il test binomiale utilizzando il primo valore trovato nel campione per definire l'"esito positivo". Questa opzione si può applicare solo ai campi nominali o ordinali con due soli valori; tutti gli altri campi categoriali specificati nella scheda Campi in cui è utilizzata questa opzione non vengono testati. È l'impostazione di default.
- Specifica valori dell'esito positivo esegue il test binomiale utilizzando l'elenco dei valori specificati per definire l'"esito positivo". Specificare un elenco di valori stringa o numerici. I valori dell'elenco non devono necessariamente essere presenti nel campione.

Definisci l'esito positivo per i campi continui. Specifica come viene definito l'"esito positivo" (il valore o i valori dei dati confrontati con il valore del test) per i campi continui. L'esito positivo viene definito come valori uguali o minori di un punto di taglio.

- Valore intermedio campione imposta il punto di taglio sulla media dei valori massimo e minimo.
- Punto di taglio personalizzato consente di specificare un valore per il punto di taglio.

Test chi-quadrato: Opzioni

Figura 27-5
Test non parametrici a campione singolo: Test chi-quadrato: Opzioni



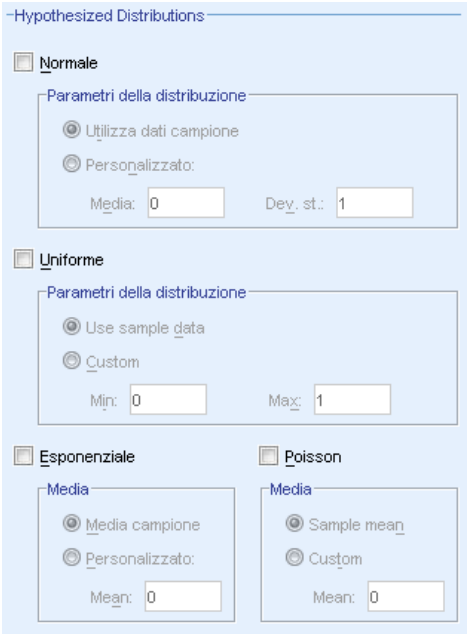
Tutte le categorie hanno uguale probabilità. Genera frequenze uguali fra tutte le categorie del campione. È l'impostazione di default.

Personalizza probabilità prevista. Consente di indicare frequenze non uguali per un elenco specificato di categorie. Specificare un elenco di valori stringa o numerici. I valori dell'elenco non devono necessariamente essere presenti nel campione. Nella colonna Categoria, specificare i valori delle categorie. Nella colonna Frequenza relativa, specificare un valore maggiore di 0 per ogni categoria. Le frequenze personalizzate vengono considerate rapporti: così, ad esempio,

specificare le frequenze 1, 2 e 3 equivale a specificare le frequenze 10, 20 e 30, ed entrambe specificano che si presume che 1/6 dei record ricada nella prima categoria, 1/3 nella seconda e 1/2 nella terza. Quando si indicano probabilità personalizzate previste, i valori delle categorie personalizzate devono includere tutti i valori del campo nei dati; in caso contrario, il test di quel campo non viene eseguito.

Opzioni test di Kolmogorov-Smirnov

Figura 27-6
Test non parametrici a campione singolo: Opzioni test di Kolmogorov-Smirnov



Questa finestra di dialogo specifica le distribuzioni da testare e i parametri delle distribuzioni ipotizzate.

Normale. Utilizza dati campione utilizza la media e la deviazione standard osservate, Personalizzato consente di specificare dei valori.

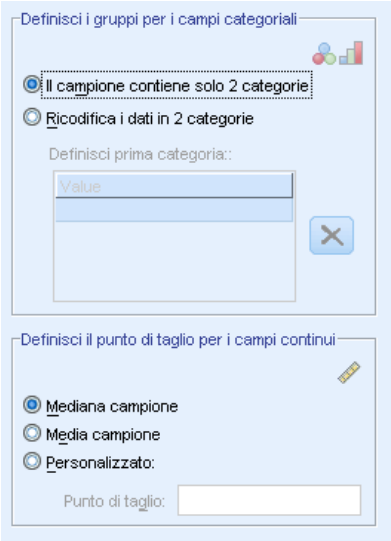
Uniforme. Utilizza dati campione utilizza il minimo e il massimo osservati, Personalizzato consente di specificare dei valori.

Esponenziale. Media campione utilizza la media osservata, Personalizzato consente di specificare dei valori.

Poisson. Media campione utilizza la media osservata, Personalizzato consente di specificare dei valori.

Test delle successioni: Opzioni

Figura 27-7
Test non parametrici a campione singolo: Test delle successioni: Opzioni



Il test delle successioni è destinato ai campi flag (campi categoriali con due sole categorie), ma può essere applicato a tutti i campi utilizzando le regole per la definizione dei gruppi.

Definisci i gruppi per i campi categoriali

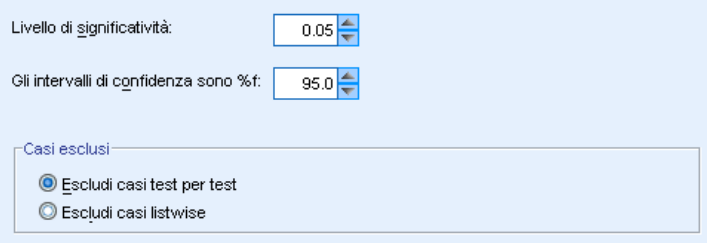
- Il campione contiene solo 2 categorie esegue il test delle successioni definendo i gruppi con i valori rilevati nel campione. Questa opzione si può applicare solo ai campi nominali o ordinali con due soli valori; tutti gli altri campi categoriali specificati nella scheda Campi in cui è utilizzata questa opzione non vengono testati.
- Ricodifica i dati in 2 categorie esegue il test delle successioni definendo uno dei gruppi con l'elenco di valori specificato. Tutti gli altri valori del campione definiscono l'altro gruppo. Non è necessario che tutti i valori dell'elenco siano presenti nel campione, ma ogni gruppo deve comprendere almeno un record.

Definisci il punto di taglio per i campi continui. Specifica come vengono definiti i gruppi per i campi continui. Il primo gruppo viene definito come valori uguali o minori di un punto di taglio.

- Mediana campione imposta il punto di taglio sulla mediana del campione.
- Media campione imposta il punto di taglio sulla media del campione.
- Personalizzato consente di specificare un valore per il punto di taglio.

Opzioni test

Figura 27-8
Test non parametrici a campione singolo: impostazioni di Opzioni test



Livello di significatività. Specifica il livello di significatività (alfa) di tutti i test. Indicare un valore numerico compreso fra 0 e 1. L'impostazione predefinita è 0,05.

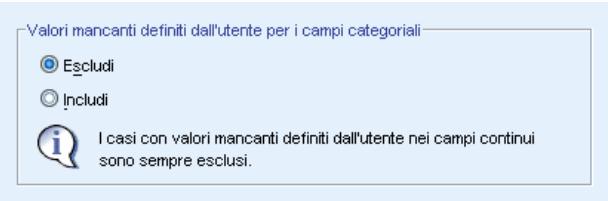
Intervallo di confidenza (%). Specifica il livello di confidenza per tutti gli intervalli di confidenza generati. Indicare un valore numerico compreso fra 0 e 100. L'impostazione predefinita è 95.

Casi esclusi. Specifica come determinare la base di casi per i test.

- Esclusione listwise significa che i record con valori mancanti in qualunque campo denominato nella scheda Campi vengono esclusi da tutte le analisi.
- Esclusione casi test per test significa che i record con valori mancanti per un campo utilizzato in un determinato test vengono omessi da quel test. Quando nell'analisi vengono specificati più test, ciascuno viene valutato separatamente.

Valori mancanti definiti dall'utente

Figura 27-9
Test non parametrici a campione singolo: impostazioni di Valori mancanti definiti dall'utente

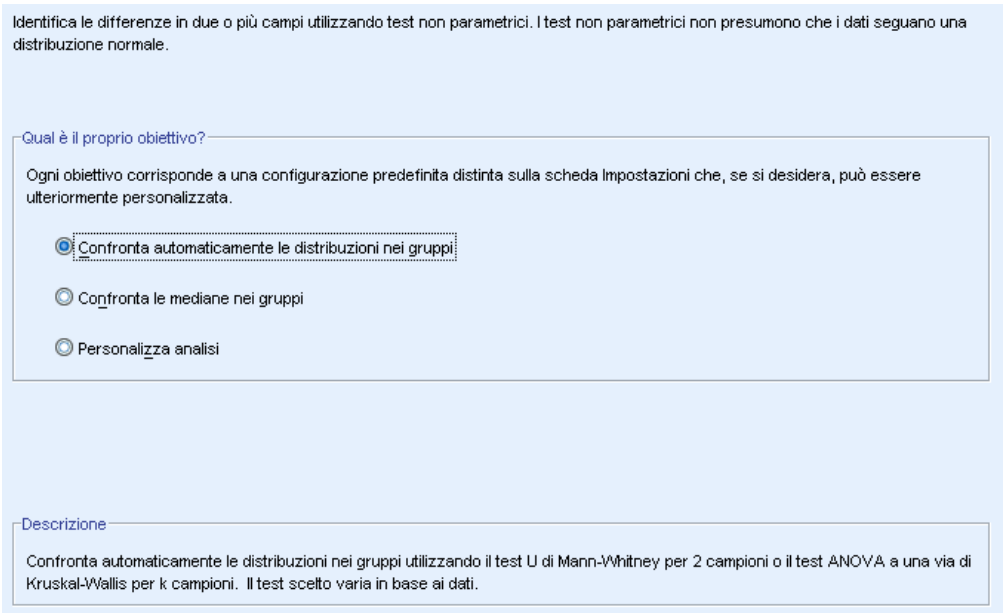


Valori mancanti definiti dall'utente per i campi categoriali. Per essere inclusi nell'analisi, i campi categoriali devono contenere valori validi per un record. Questi controlli consentono di decidere se i valori mancanti definiti dall'utente devono essere considerati validi nell'ambito dei campi categoriali. I valori mancanti di sistema e i valori mancanti relativi ai campi continui vengono sempre considerati non validi.

Test non parametrici a campioni indipendenti

I test non parametrici a campioni indipendenti individuano le differenze fra due o più gruppi mediante uno o più test non parametrici. I test non parametrici non presumono che i dati seguano una distribuzione normale.

Figura 27-10
Test non parametrici a campioni indipendenti: scheda Obiettivo



Qual è il proprio obiettivo? Gli obiettivi consentono di specificare rapidamente varie impostazioni di uso comune per i test.

- **Confronta automaticamente le distribuzioni nei gruppi.** Questo obiettivo applica il test U di Mann-Whitney ai dati con 2 gruppi o il test ANOVA a una via di Kruskal-Wallis ai dati con *k* gruppi.
- **Confronta le mediane nei gruppi.** Questo obiettivo utilizza il test della mediana per confrontare le mediane osservate tra più gruppi.
- **Analisi personalizzata.** Selezionare questa opzione se si desidera modificare manualmente le impostazioni del test nella scheda Impostazioni. Si noti che questa impostazione viene selezionata automaticamente se in seguito si apportano modifiche incompatibili con l'obiettivo selezionato alle opzioni della scheda Impostazioni.

Per ottenere test non parametrici a campioni indipendenti

Dai menu, scegliere:
Analizza > Test non parametrici > Campioni indipendenti...

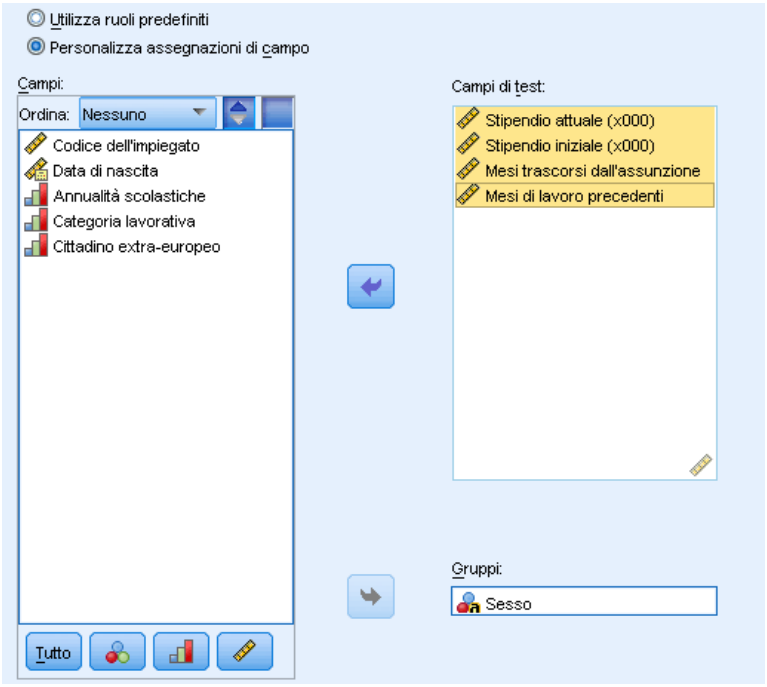
▶ Fare clic su Esegui.

Se lo si desidera, è possibile:

- Specificare un obiettivo nella scheda Obiettivi.
- Specificare le assegnazioni di campo nella scheda Campi.
- Specificare delle impostazioni avanzate nella scheda Impostazioni.

Scheda Campi

Figura 27-11
Test non parametrici a campioni indipendenti: scheda Campi



La scheda Campi indica i campi da testare e il campo utilizzato per definire i gruppi.

Utilizza ruoli predefiniti. Questa opzione utilizza le informazioni contenute nei campi esistenti. Tutti i campi continui con un ruolo predefinito come Obiettivo o Entrambi saranno utilizzati come campi di test. Se esiste un solo campo categoriale con un ruolo predefinito come Input viene utilizzato come campo di raggruppamento. In caso contrario, non viene utilizzato alcun campo di raggruppamento per impostazione predefinita ed è necessario utilizzare le assegnazioni campi personalizzate. È obbligatorio disporre di almeno un campo di test e di un campo di raggruppamento.

Utilizza assegnazioni campi personalizzate. Questa opzione consente di ignorare i ruoli dei campi. Dopo averla selezionata, compilare i campi riportati sotto:

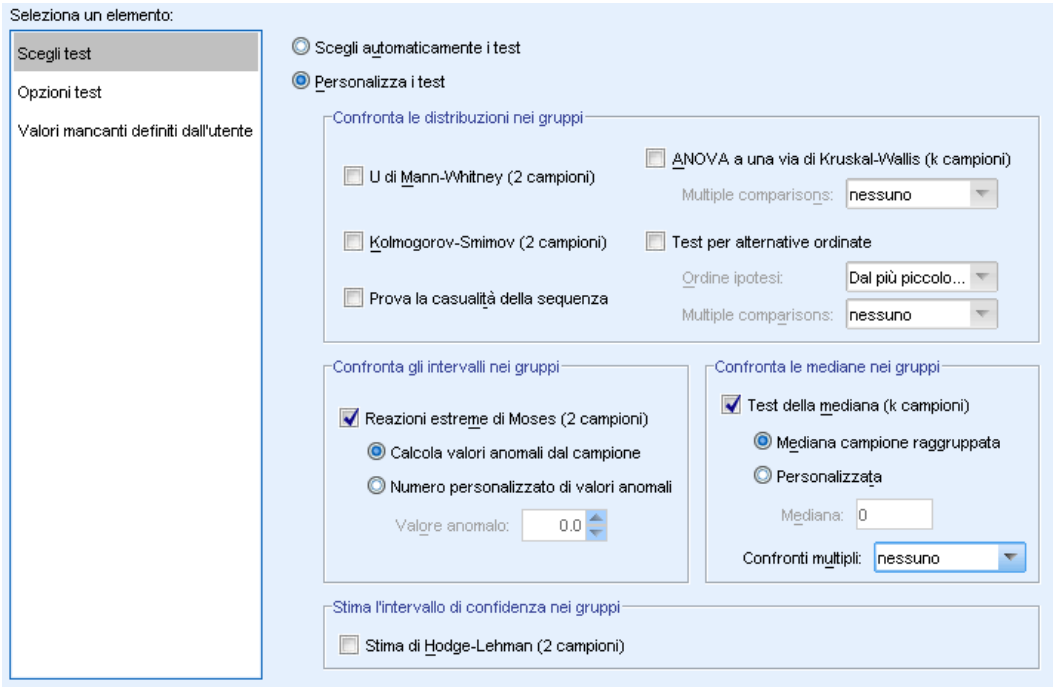
- **Campi di test.** Selezionare uno o più campi continui.
- **Gruppi.** Selezionare un campo categoriale.

Scheda Impostazioni

La scheda Impostazioni contiene vari gruppi di impostazioni che è possibile modificare per perfezionare l'elaborazione dei dati da parte dell'algoritmo. Se si apportano modifiche alle impostazioni predefinite che risultano incompatibili con l'obiettivo selezionato, la scheda Obiettivo viene aggiornata automaticamente per selezionare l'opzione Personalizza analisi.

Scegli test

Figura 27-12
Test non parametrici a campioni indipendenti: impostazioni di Scegli test



Queste impostazioni specificano i test da eseguire sui campi indicati nella scheda Campi.

Scegli automaticamente i test. Questa impostazione applica il test U di Mann-Whitney ai dati con 2 gruppi o il test ANOVA a una via di Kruskal-Wallis ai dati con *k* gruppi.

Personalizza i test. Questa impostazione consente di definire l'esecuzione di test specifici.

- **Confronta le distribuzioni nei gruppi.** Questa impostazione genera test a campioni indipendenti per verificare se i campioni appartengono alla stessa popolazione.

  U di Mann-Whitney (2 campioni) utilizza il rango di ogni caso per verificare se i gruppi sono estratti dalla stessa popolazione. Il primo valore in ordine crescente del campo di raggruppamento definisce il primo gruppo, mentre il secondo definisce il secondo gruppo. Se il campo di raggruppamento ha più di due valori, il test non viene generato.

  Kolmogorov-Smirnov (2 campioni) è sensibile a tutte le differenze di mediana, dispersione, asimmetria e simili fra le due distribuzioni. Se il campo di raggruppamento ha più di due valori, il test non viene generato.

  Sequenza di test per casualità (Wald-Wolfowitz per 2 campioni) genera un test delle successioni secondo il criterio dell'appartenenza a un gruppo. Se il campo di raggruppamento ha più di due valori, il test non viene generato.

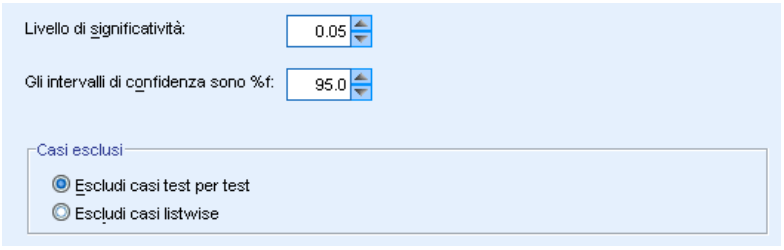
ANOVA a 1 via di Kruskal-Wallis (k campioni) è un'ampliamento del test U di Mann-Whitney ed è l'analogo non parametrico dell'analisi della varianza a una via. Se lo si desidera è possibile richiedere confronti multipli dei *k* campioni, ovvero confronti multipli tutto per coppie o confronti stepwise decrescente.

Test per alternative ordinate (Jonckheere-Terpstra per k campioni) è un'alternativa più potente al test di Kruskal-Wallis quando i *k* campioni hanno un ordinamento naturale. Ad esempio, le *k* popolazioni possono rappresentare *k* temperature crescenti. L'ipotesi che diverse temperature producano la stessa distribuzione della risposta è verificata rispetto all'ipotesi alternativa in base a cui al salire della temperatura, cresce il valore della risposta. Qui l'ipotesi alternativa è ordinata e quindi il test di Jonckheere-Terpstra è il più appropriato da utilizzare. Specificare l'ordine delle ipotesi alternative; Dal più piccolo al più grande stabilisce un'ipotesi alternativa secondo cui il parametro di posizione del primo gruppo non è uguale al secondo, che a sua volta non è uguale al terzo e così via; Dal più grande al più piccolo stabilisce un'ipotesi alternativa secondo cui il parametro di posizione dell'ultimo gruppo non è uguale al penultimo, che a sua volta non è uguale al terzultimo e così via. Se lo si desidera è possibile richiedere confronti multipli dei *k* campioni, ovvero confronti multipli tutto per coppie o confronti stepwise decrescente.

- **Confronta gli intervalli nei gruppi.** Questa opzione genera un test a campioni indipendenti per verificare se i campioni hanno lo stesso intervallo. Reazioni estreme di Moses (2 campioni) verifica un gruppo di controllo con un gruppo di confronto. Il primo valore in ordine crescente del campo di raggruppamento definisce il gruppo di controllo, mentre il secondo definisce il gruppo di confronto. Se il campo di raggruppamento ha più di due valori, il test non viene generato.
- **Confronta le mediane nei gruppi.** Questa opzione genera un test a campioni indipendenti per verificare se i campioni hanno la stessa mediana. Test della mediana (k campioni) può utilizzare come mediana ipotizzata sia la mediana campione raggruppata (calcolata su tutti i record dell'insieme di dati), sia un valore personalizzato. Se lo si desidera è possibile richiedere confronti multipli dei *k* campioni, ovvero confronti multipli tutto per coppie o confronti stepwise decrescente.
- **Stima l'intervallo di confidenza nei gruppi.** Stima di Hodges-Lehman (2 campioni) genera una stima a campioni indipendenti e un intervallo di confidenza per la differenza tra le mediane di due gruppi. Se il campo di raggruppamento ha più di due valori, il test non viene generato.

Opzioni test

Figura 27-13
Test non parametrici a campioni indipendenti: impostazioni di Opzioni test



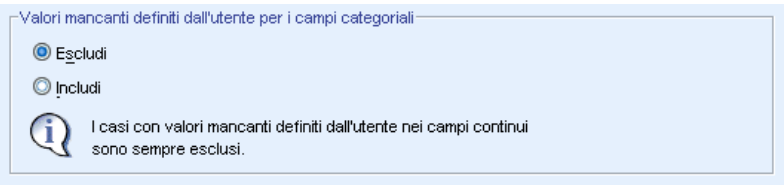
Livello di significatività. Specifica il livello di significatività (alfa) di tutti i test. Indicare un valore numerico compreso fra 0 e 1. L'impostazione predefinita è 0,05.

Intervallo di confidenza (%). Specifica il livello di confidenza per tutti gli intervalli di confidenza generati. Indicare un valore numerico compreso fra 0 e 100. L'impostazione predefinita è 95.

Casi esclusi. Specifica come determinare la base di casi per i test. Esclusione listwise significa che i record con valori mancanti in qualunque campo denominato in un sottocomando vengono esclusi da tutte le analisi. Esclusione casi test per test significa che i record con valori mancanti per un campo utilizzato in un determinato test vengono omessi da quel test. Quando nell'analisi vengono specificati più test, ciascuno viene valutato separatamente.

Valori mancanti definiti dall'utente

Figura 27-14
Test non parametrici a campioni indipendenti: impostazioni di Valori mancanti definiti dall'utente



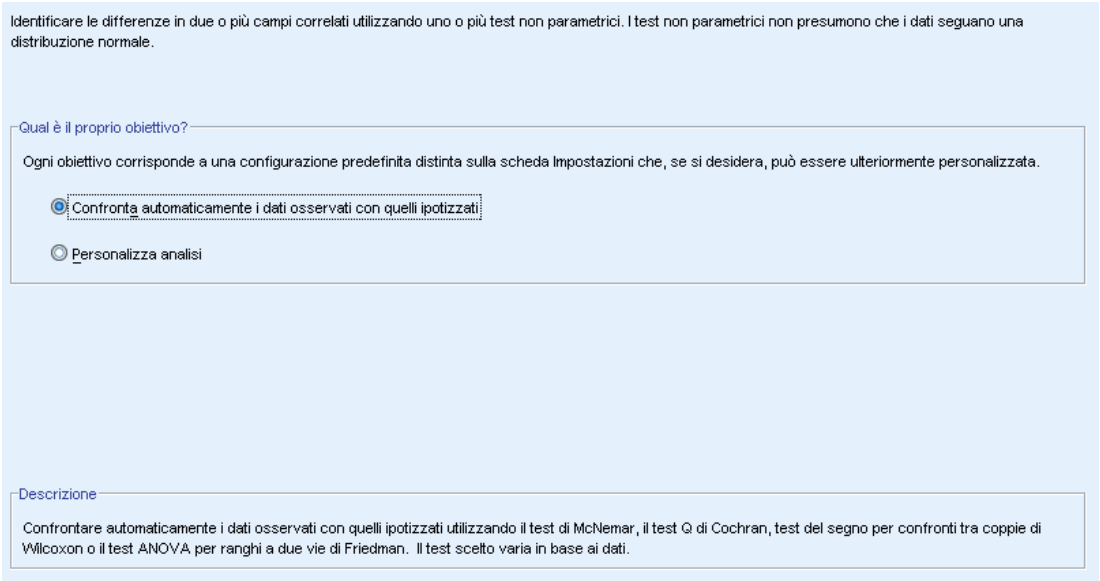
Valori mancanti definiti dall'utente per i campi categoriali. Per essere inclusi nell'analisi, i campi categoriali devono contenere valori validi per un record. Questi controlli consentono di decidere se i valori mancanti definiti dall'utente devono essere considerati validi nell'ambito dei campi categoriali. I valori mancanti di sistema e i valori mancanti relativi ai campi continui vengono sempre considerati non validi.

Test non parametrici a campioni correlati

Identificano le differenze fra due o più campi correlati utilizzando uno o più test non parametrici. I test non parametrici non presumono che i dati seguano una distribuzione normale.

Considerazioni sui dati. Ogni record corrisponde a un determinato soggetto per cui vengono archiviate due o più misurazioni correlate in campi separati dell'insieme di dati. Per esempio, uno studio relativo all'efficacia di un regime alimentare può essere analizzato mediante test non parametrici a campioni correlati se il peso di ogni soggetto è misurato a intervalli regolari e memorizzato in campi come *Peso prima della dieta*, *Peso durante la dieta* e *Peso dopo la dieta*. Questi campi sono "correlati".

Figura 27-15
Test non parametrici a campioni correlati: scheda Obiettivo



Qual è il proprio obiettivo? Gli obiettivi consentono di specificare rapidamente varie impostazioni di uso comune per i test.

- **Confronta automaticamente i dati osservati con quelli ipotizzati.** Questo obiettivo applica ai dati categoriali il test di McNemar quando vengono specificati 2 campi e il test Q di Cochran quando vengono specificati più di 2 campi; ai dati continui, il test del segno per confronti tra coppie di Wilcoxon quando vengono specificati 2 campi e il test ANOVA per ranghi a due vie di Friedman quando vengono specificati più di 2 campi.
- **Analisi personalizzata.** Selezionare questa opzione se si desidera modificare manualmente le impostazioni del test nella scheda Impostazioni. Si noti che questa impostazione viene selezionata automaticamente se in seguito si apportano modifiche incompatibili con l'obiettivo selezionato alle opzioni della scheda Impostazioni.

Quando si specificano campi con livelli di misurazione diversi, essi vengono prima separati in base al livello di misurazione e quindi a ogni gruppo viene applicato il test appropriato. Per esempio, se si sceglie come obiettivo Confronta automaticamente i dati osservati con quelli ipotizzati e si specificano 3 campi continui e 2 campi nominali, ai campi continui viene applicato il test di Friedman e a quelli nominali viene applicato il test di McNemar.

Per ottenere test non parametrici a campioni correlati

Dai menu, scegliere:
Analizza > Test non parametrici > Campioni correlati...

► Fare clic su Esegui.

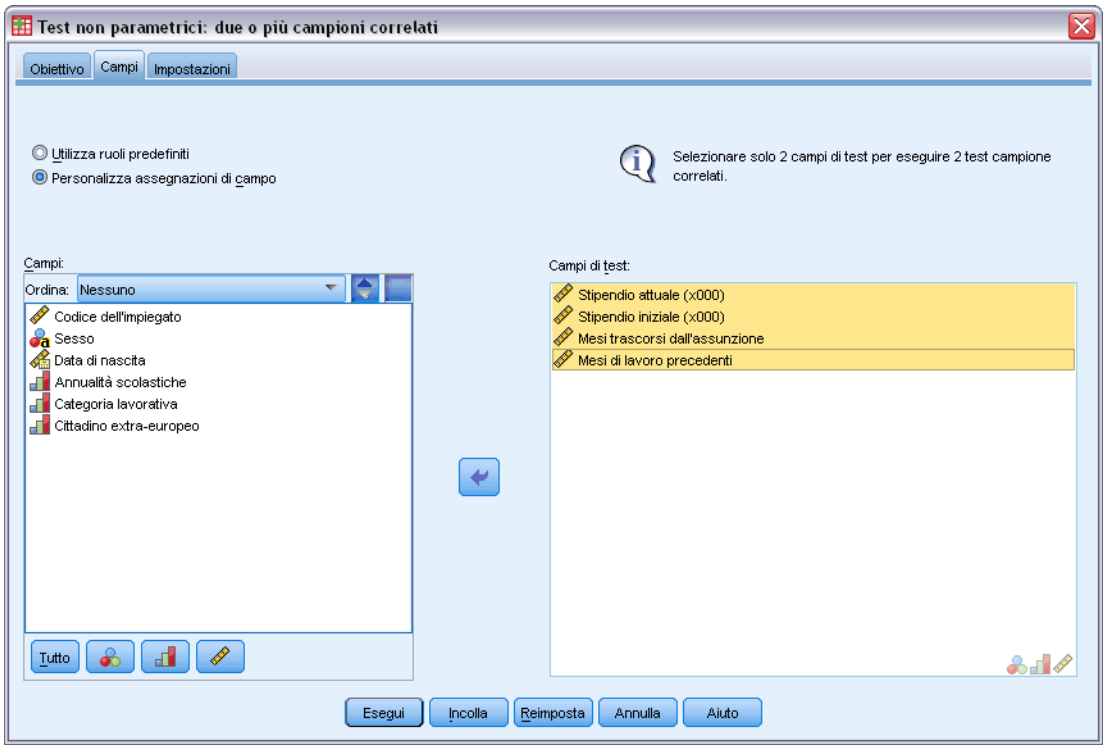
Se lo si desidera, è possibile:

- Specificare un obiettivo nella scheda Obiettivi.

- Specificare le assegnazioni di campo nella scheda Campi.
- Specificare delle impostazioni avanzate nella scheda Impostazioni.

Scheda Campi

Figura 27-16
Test non parametrici a campioni correlati: scheda Campi



La scheda Campi indica i campi che è necessario testare.

Utilizza ruoli predefiniti. Questa opzione utilizza le informazioni contenute nei campi esistenti. Tutti i campi con un ruolo predefinito come Obiettivo o Entrambi saranno utilizzati come campi di test. È obbligatorio avere almeno due campi di test.

Utilizza assegnazioni campi personalizzate. Questa opzione consente di ignorare i ruoli dei campi. Dopo averla selezionata, compilare i campi riportati sotto:

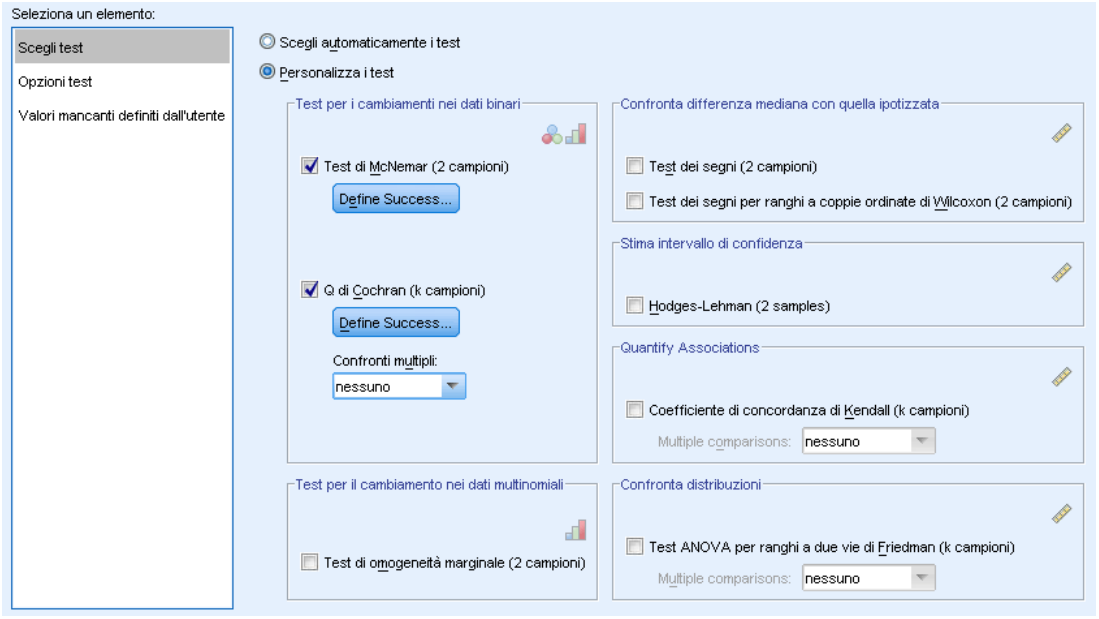
- **Campi di test.** Selezionare due o più campi. Ogni campo rappresenta un campione correlato diverso.

Scheda Impostazioni

La scheda Impostazioni contiene vari gruppi di impostazioni che è possibile modificare per perfezionare l'elaborazione dei dati da parte della procedura. Se si apportano modifiche alle impostazioni predefinite che risultano incompatibili con gli altri obiettivi, la scheda Obiettivo viene aggiornata automaticamente per selezionare l'opzione Personalizza analisi.

Scegli test

Figura 27-17
Test non parametrici a campioni correlati: impostazioni di Scegli test



Queste impostazioni specificano i test da eseguire sui campi indicati nella scheda Campi.

Scegli automaticamente i test. Questa impostazione applica ai dati categoriali il test di McNemar quando vengono specificati 2 campi e il test Q di Cochran quando vengono specificati più di 2 campi; ai dati continui, il test del segno per confronti tra coppie di Wilcoxon quando vengono specificati 2 campi e il test ANOVA per ranghi a due vie di Friedman quando vengono specificati più di 2 campi.

Personalizza i test. Questa impostazione consente di definire l'esecuzione di test specifici.

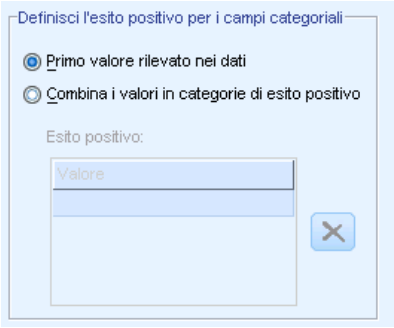
- **Test per i cambiamenti nei dati binari.** Test di McNemar (2 campioni) si può applicare ai campi categoriali. Questa opzione genera un test a campioni correlati per verificare se le combinazioni di valori tra due campi flag (campi categoriali con due soli valori) hanno uguale probabilità. Se nella scheda Campi sono specificati più di due campi, il test non viene eseguito. Vedere Test di McNemar: Definisci esito positivo per informazioni più dettagliate sulle impostazioni dei test. Q di Cochran (k campioni) si può applicare ai campi categoriali. Questa opzione genera un test a campioni correlati per verificare se le combinazioni di valori tra *k* campi flag (campi categoriali con due soli valori) hanno uguale probabilità. Se lo si desidera è possibile richiedere confronti multipli dei *k* campioni, ovvero confronti multipli tutto per coppie o confronti stepwise decrescente. Vedere Q di Cochran: Definisci esito positivo per informazioni più dettagliate sulle impostazioni dei test.
- **Test per il cambiamento nei dati multinomiali.** Test di omogeneità marginale (2 campioni) genera un test a campioni correlati per verificare se le combinazioni di valori tra due campi ordinali appaiati hanno uguale probabilità. Il test di omogeneità marginale si utilizza in genere nelle situazioni in cui sono presenti misure ripetute. Estensione del test di McNemar dalla risposta

binaria a quella multinomiale. Se nella scheda Campi sono specificati più di due campi, il test non viene eseguito.

- **Confronta differenza mediana con quella ipotizzata.** Ciascuno di questi test genera un test a campioni correlati per verificare se la differenza mediana fra due campi continui è diversa da 0. Se nella scheda Campi sono specificati più di due campi, i test non vengono eseguiti.
- **Stima intervallo di confidenza.** Genera una stima a campioni correlati e un intervallo di confidenza per la differenza mediana fra due campi continui appaiati. Se nella scheda Campi sono specificati più di due campi, il test non viene eseguito.
- **Quantifica associazioni.** Coefficiente di concordanza di Kendall (k campioni) genera una misura di accordo tra giudici o stimatori in cui ogni record rappresenta la valutazione di vari elementi (campi) da parte di un giudice. Se lo si desidera è possibile richiedere confronti multipli dei *k* campioni, ovvero confronti multipli tutto per coppie o confronti stepwise decrescente.
- **Confronta distribuzioni.** Test ANOVA per ranghi a due vie di Friedman (k campioni) genera un test a campioni correlati per verificare se *k* campioni correlati sono stati estratti dalla stessa popolazione. Se lo si desidera è possibile richiedere confronti multipli dei *k* campioni, ovvero confronti multipli tutto per coppie o confronti stepwise decrescente.

Test di McNemar: Definisci esito positivo

Figura 27-18
Test non parametrici a campioni correlati: Test di McNemar: impostazioni di Definisci esito positivo



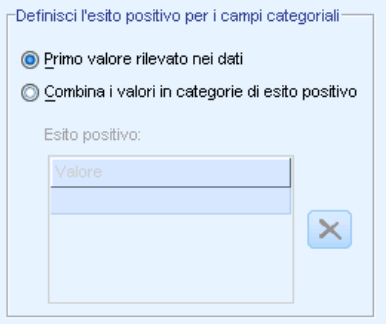
Il test di McNemar è destinato ai campi flag (campi categoriali con due sole categorie), ma viene applicato a tutti i campi categoriali utilizzando le regole per la definizione dell'"esito positivo".

Definisci l'esito positivo per i campi categoriali. Specifica come viene definito l'"esito positivo" per i campi categoriali.

- Utilizza la prima categoria trovata nei dati esegue il test utilizzando il primo valore trovato nel campione per definire l'"esito positivo". Questa opzione si può applicare solo ai campi nominali o ordinali con due soli valori; tutti gli altri campi categoriali specificati nella scheda Campi in cui è utilizzata questa opzione non vengono testati. È l'impostazione predefinita.
- Specifica valori dell'esito positivo esegue il test utilizzando l'elenco dei valori specificati per definire l'"esito positivo". Specificare un elenco di valori stringa o numerici. I valori dell'elenco non devono necessariamente essere presenti nel campione.

Q di Cochran: Definisci esito positivo

Figura 27-19
Test non parametrici a campioni correlati: Test Q di Cochran: Definisci esito positivo



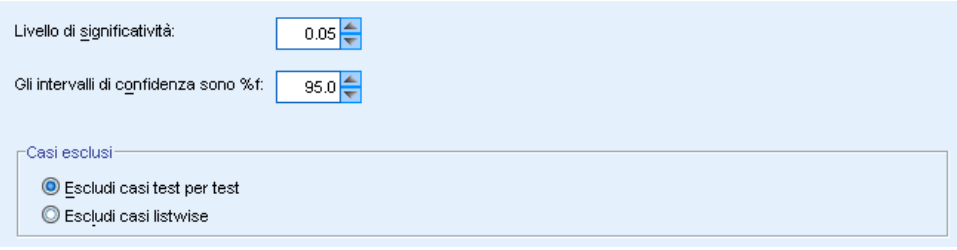
Il test Q di Cochran è destinato ai campi flag (campi categoriali con due sole categorie), ma viene applicato a tutti i campi categoriali utilizzando le regole per la definizione dell'"esito positivo".

Definisci l'esito positivo per i campi categoriali. Specifica come viene definito l'"esito positivo" per i campi categoriali.

- Utilizza la prima categoria trovata nei dati esegue il test utilizzando il primo valore trovato nel campione per definire l'"esito positivo". Questa opzione si può applicare solo ai campi nominali o ordinali con due soli valori; tutti gli altri campi categoriali specificati nella scheda Campi in cui è utilizzata questa opzione non vengono testati. È l'impostazione predefinita.
- Specifica valori dell'esito positivo esegue il test utilizzando l'elenco dei valori specificati per definire l'"esito positivo". Specificare un elenco di valori stringa o numerici. I valori dell'elenco non devono necessariamente essere presenti nel campione.

Opzioni test

Figura 27-20
Test non parametrici a campioni correlati: impostazioni di Opzioni test



Livello di significatività. Specifica il livello di significatività (alfa) di tutti i test. Indicare un valore numerico compreso fra 0 e 1. L'impostazione predefinita è 0,05.

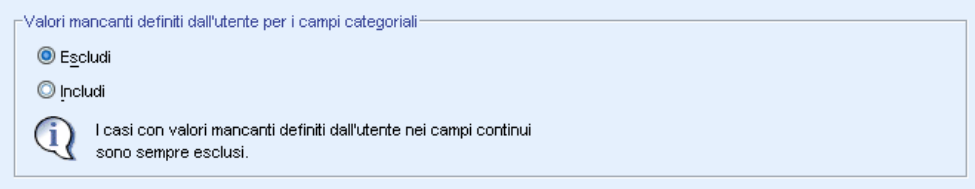
Intervallo di confidenza (%). Specifica il livello di confidenza per tutti gli intervalli di confidenza generati. Indicare un valore numerico compreso fra 0 e 100. L'impostazione predefinita è 95.

Casi esclusi. Specifica come determinare la base di casi per i test.

- Esclusione listwise significa che i record con valori mancanti in qualunque campo denominato in un sottocomando vengono esclusi da tutte le analisi.
- Esclusione casi test per test significa che i record con valori mancanti per un campo utilizzato in un determinato test vengono omessi da quel test. Quando nell'analisi vengono specificati più test, ciascuno viene valutato separatamente.

Valori mancanti definiti dall'utente

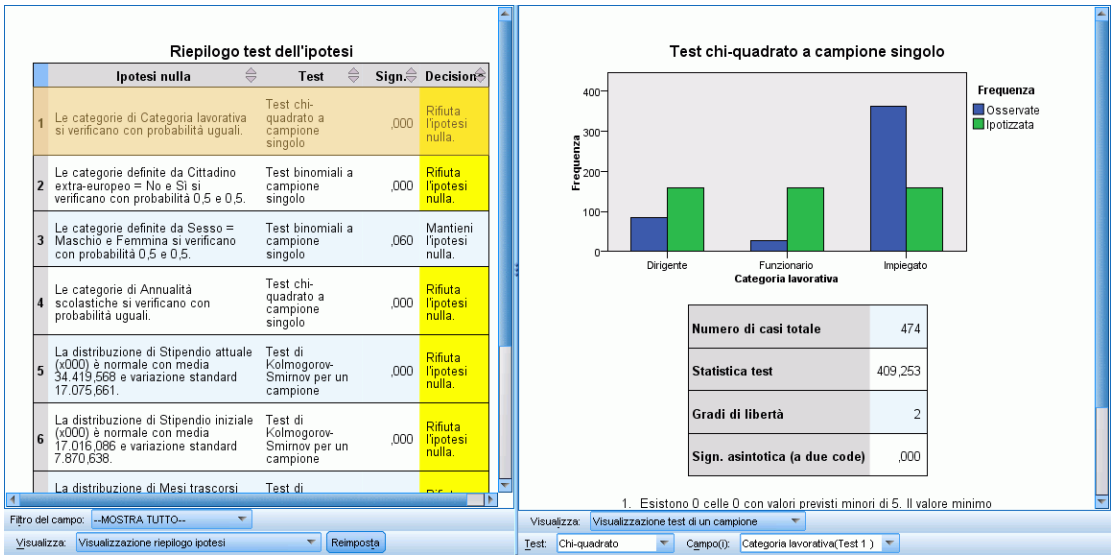
Figura 27-21
Test non parametrici a campioni correlati: impostazioni di Valori mancanti definiti dall'utente



Valori mancanti definiti dall'utente per i campi categoriali. Per essere inclusi nell'analisi, i campi categoriali devono contenere valori validi per un record. Questi controlli consentono di decidere se i valori mancanti definiti dall'utente devono essere considerati validi nell'ambito dei campi categoriali. I valori mancanti di sistema e i valori mancanti relativi ai campi continui vengono sempre considerati non validi.

Vista del modello

Figura 27-22
Vista del modello Test non parametrici



Questa procedura crea un oggetto Viewer modelli nel Viewer. Attivando l'oggetto con un doppio clic, si accede a una vista interattiva del modello La vista del modello è composta da una finestra a due riquadri, la visualizzazione principale a sinistra e quella collegata o ausiliaria a destra.

Le visualizzazioni principali sono due:

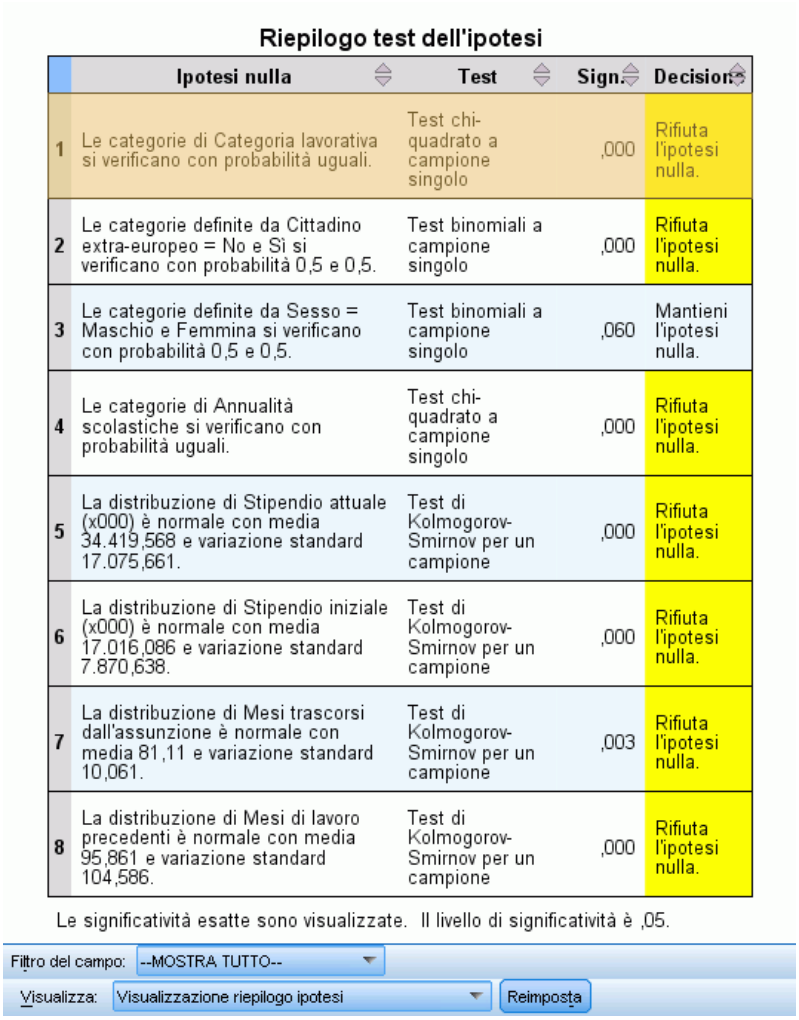
- Riepilogo ipotesi. È la visualizzazione predefinita.Per ulteriori informazioni, vedere l'argomento Riepilogo ipotesi a pag. 216.
- Riepilogo intervallo di confidenza. Per ulteriori informazioni, vedere l'argomento Riepilogo intervallo di confidenza a pag. 217.

Le visualizzazioni collegate/ausiliarie sono sette:

- Test di un campione. Questa è la visualizzazione predefinita quando si richiedono test a campione singolo. Per ulteriori informazioni, vedere l'argomento Test di un campione a pag. 218.
- Test campioni correlati. Questa è la visualizzazione predefinita quando si richiedono solo test a campioni correlati e non test a campione singolo. Per ulteriori informazioni, vedere l'argomento Test campioni correlati a pag. 222.
- Test campioni indipendenti. Questa è la visualizzazione predefinita quando non si richiedono test a campioni correlati né a campione singolo. Per ulteriori informazioni, vedere l'argomento Test campioni indipendenti a pag. 229.
- Informazioni sul campo categoriale. Per ulteriori informazioni, vedere l'argomento Informazioni sul campo categoriale a pag. 237.
- Informazioni sul campo continuo. Per ulteriori informazioni, vedere l'argomento Informazioni sul campo continuo a pag. 238.
- Confronti pairwise. Per ulteriori informazioni, vedere l'argomento Confronti pairwise a pag. 239.
- Sottoinsiemi omogenei. Per ulteriori informazioni, vedere l'argomento Sottoinsiemi omogenei a pag. 240.

Riepilogo ipotesi

Figura 27-23
Riepilogo ipotesi



Riepilogo test dell'ipotesi

	Ipotesi nulla	Test	Sign.	Decision
1	Le categorie di Categoria lavorativa si verificano con probabilità uguali.	Test chi-quadrato a campione singolo	,000	Rifiuta l'ipotesi nulla.
2	Le categorie definite da Cittadino extra-europeo = No e Sì si verificano con probabilità 0,5 e 0,5.	Test binomiali a campione singolo	,000	Rifiuta l'ipotesi nulla.
3	Le categorie definite da Sesso = Maschio e Femmina si verificano con probabilità 0,5 e 0,5.	Test binomiali a campione singolo	,060	Mantieni l'ipotesi nulla.
4	Le categorie di Annualità scolastiche si verificano con probabilità uguali.	Test chi-quadrato a campione singolo	,000	Rifiuta l'ipotesi nulla.
5	La distribuzione di Stipendio attuale (x000) è normale con media 34.419,568 e variazione standard 17.075,661.	Test di Kolmogorov-Smirnov per un campione	,000	Rifiuta l'ipotesi nulla.
6	La distribuzione di Stipendio iniziale (x000) è normale con media 17.016,086 e variazione standard 7.870,638.	Test di Kolmogorov-Smirnov per un campione	,000	Rifiuta l'ipotesi nulla.
7	La distribuzione di Mesi trascorsi dall'assunzione è normale con media 81,11 e variazione standard 10,061.	Test di Kolmogorov-Smirnov per un campione	,003	Rifiuta l'ipotesi nulla.
8	La distribuzione di Mesi di lavoro precedenti è normale con media 95,861 e variazione standard 104,586.	Test di Kolmogorov-Smirnov per un campione	,000	Rifiuta l'ipotesi nulla.

Le significatività esatte sono visualizzate. Il livello di significatività è ,05.

Filtro del campo: --MOSTRA TUTTO--

Visualizza: Visualizzazione riepilogo ipotesi Reimposta

La visualizzazione Riepilogo del modello è un'istantanea, un riepilogo immediato dei test non parametrici. Essa evidenzia le ipotesi e le decisioni nulle, concentrando l'attenzione sui valori *p* significativi.

- Ogni riga corrisponde a un test diverso. Fare clic su una riga per visualizzare ulteriori informazioni sul test nella visualizzazione collegata.
- Fare clic sull'intestazione di una colonna per ordinare le righe in base ai valori di quella colonna.
- Il pulsante di ripristino consente di reimpostare il Viewer modelli sullo stato originario.
- L'elenco a discesa Filtro del campo consente di visualizzare solo i test che riguardano il campo selezionato. Ad esempio, quando si seleziona *Stipendio iniziale* nell'elenco a discesa Filtro del campo, nel Riepilogo ipotesi vengono visualizzati solo due test.

Figura 27-24
Riepilogo ipotesi dopo l'applicazione del filtro Stipendio iniziale

Riepilogo test dell'ipotesi

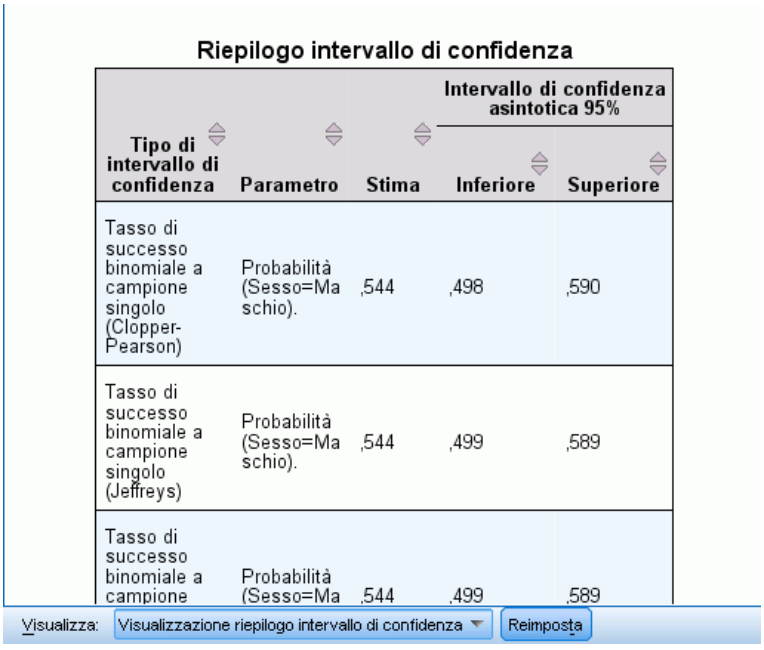
	Ipotesi nulla	Test	Sign.	Decisione
6	La distribuzione di Stipendio iniziale (x000) è normale con media 17.016,086 e variazione standard 7.870,638.	Test di Kolmogorov-Smirnov per un campione	,000	Rifiuta l'ipotesi nulla.

Le significatività esatte sono visualizzate. Il livello di significatività è ,05.

Filtro del campo: Stipendio iniziale (x000)
Visualizza: Visualizzazione riepilogo ipotesi Reimposta

Riepilogo intervallo di confidenza

Figura 27-25
Riepilogo intervallo di confidenza



Riepilogo intervallo di confidenza

Tipo di intervallo di confidenza	Parametro	Stima	Intervallo di confidenza asintotica 95%	
			Inferiore	Superiore
Tasso di successo binomiale a campione singolo (Clopper-Pearson)	Probabilità (Sesso=Maschio).	,544	,498	,590
Tasso di successo binomiale a campione singolo (Jeffreys)	Probabilità (Sesso=Maschio).	,544	,499	,589
Tasso di successo binomiale a campione	Probabilità (Sesso=Ma	,544	,499	,589

Visualizza: Visualizzazione riepilogo intervallo di confidenza Reimposta

Il Riepilogo intervallo di confidenza mostra gli intervalli di confidenza generati dai test non parametrici.

- Ogni riga corrisponde a un intervallo di confidenza diverso.
- Fare clic sull'intestazione di una colonna per ordinare le righe in base ai valori di quella colonna.

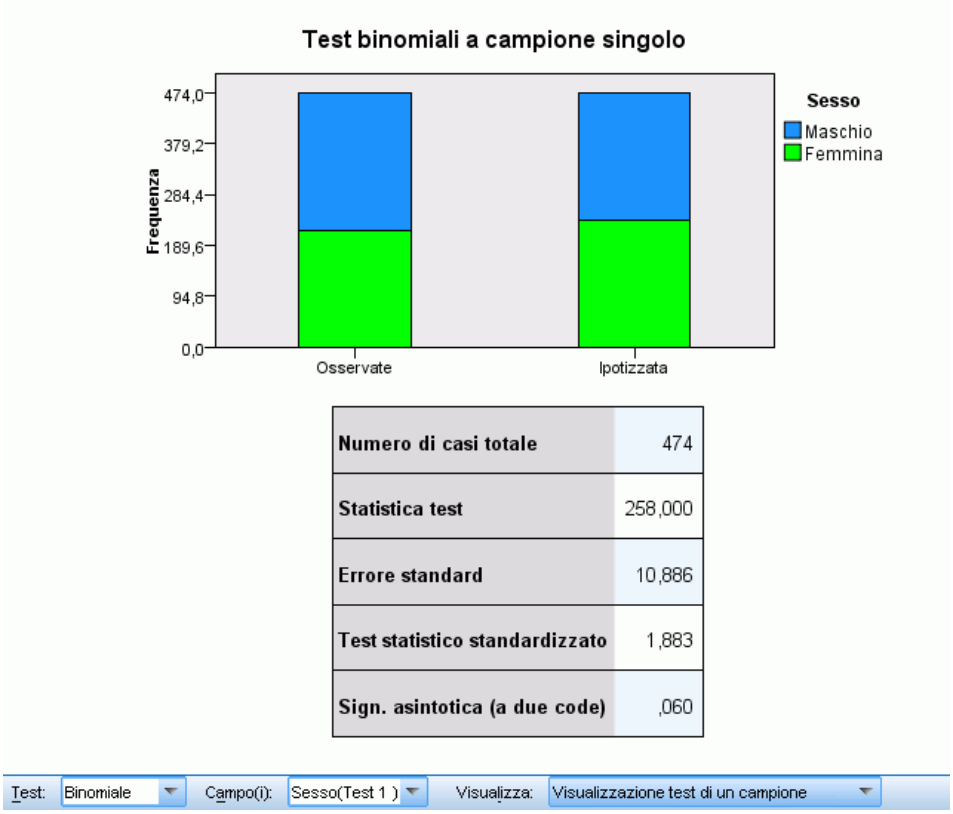
Test di un campione

La Visualizzazione test di un campione mostra i dettagli relativi a tutti i test non parametrici a campione singolo richiesti. Le informazioni visualizzate dipendono dal test selezionato.

- L'elenco a discesa Test consente di selezionare il tipo di test a campione singolo desiderato.
- L'elenco a discesa Campo(i) consente di selezionare un campo sottoposto al test selezionato nell'elenco a discesa Test.

Test binomiale

Figura 27-26
Visualizzazione test di un campione, Test binomiale

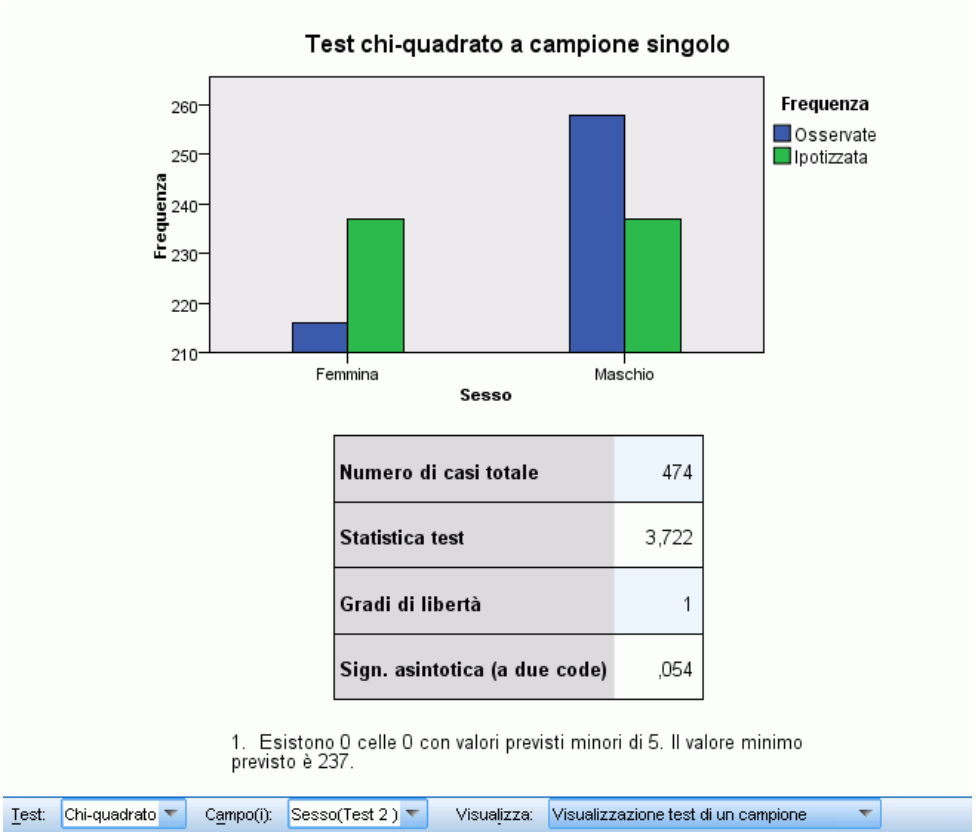


Il Test binomiale mostra un grafico a barre sovrapposto e una tabella dei test.

- Il grafico a barre sovrapposto mostra le frequenze osservate e ipotizzate per le categorie "esito positivo" ed "esito negativo" del campo di test, con gli "esiti negativi" sovrapposti agli "esiti positivi". Se si passa il mouse sopra una barra viene visualizzata una descrizione con le percentuali della categoria. Differenze visibili tra le barre indicano che il campo di test può non presentare la distribuzione binomiale ipotizzata.
- La tabella mostra i dettagli del test.

Test Chi-quadrato

Figura 27-27
Visualizzazione test di un campione, Test chi-quadrato



Numero di casi totale	474
Statistica test	3,722
Gradi di libertà	1
Sign. asintotica (a due code)	,054

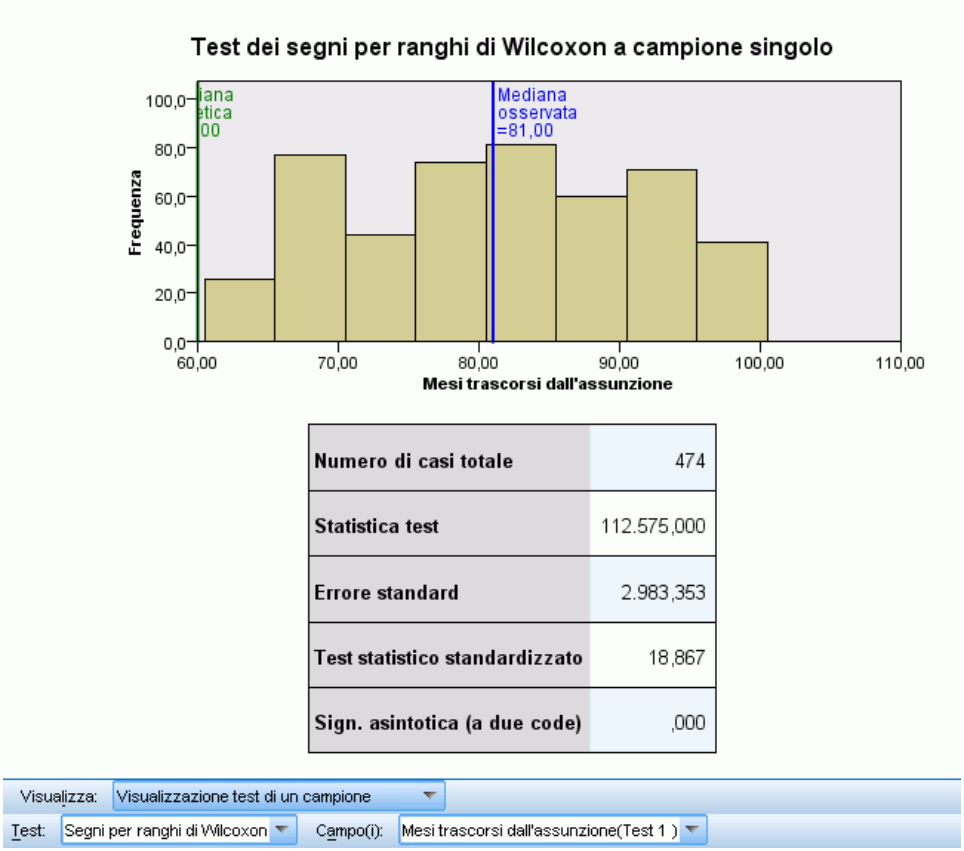
1. Esistono 0 celle 0 con valori previsti minori di 5. Il valore minimo previsto è 237.

La visualizzazione Test Chi-quadrato mostra un grafico a barre raggruppato e una tabella dei test.

- Il grafico a barre raggruppato mostra le frequenze osservate e ipotizzate per ogni categoria del campo di test. Se si passa il mouse sopra una barra viene visualizzata una descrizione delle frequenze osservate e ipotizzate e la relativa differenza (residuo). Differenze visibili tra le barre corrispondenti alla frequenza osservata e a quella ipotizzata indicano che il campo di test può non presentare la distribuzione ipotizzata.
- La tabella mostra i dettagli del test.

Segni per ranghi di Wilcoxon

Figura 27-28
Visualizzazione test di un campione, Test dei segni per ranghi di Wilcoxon

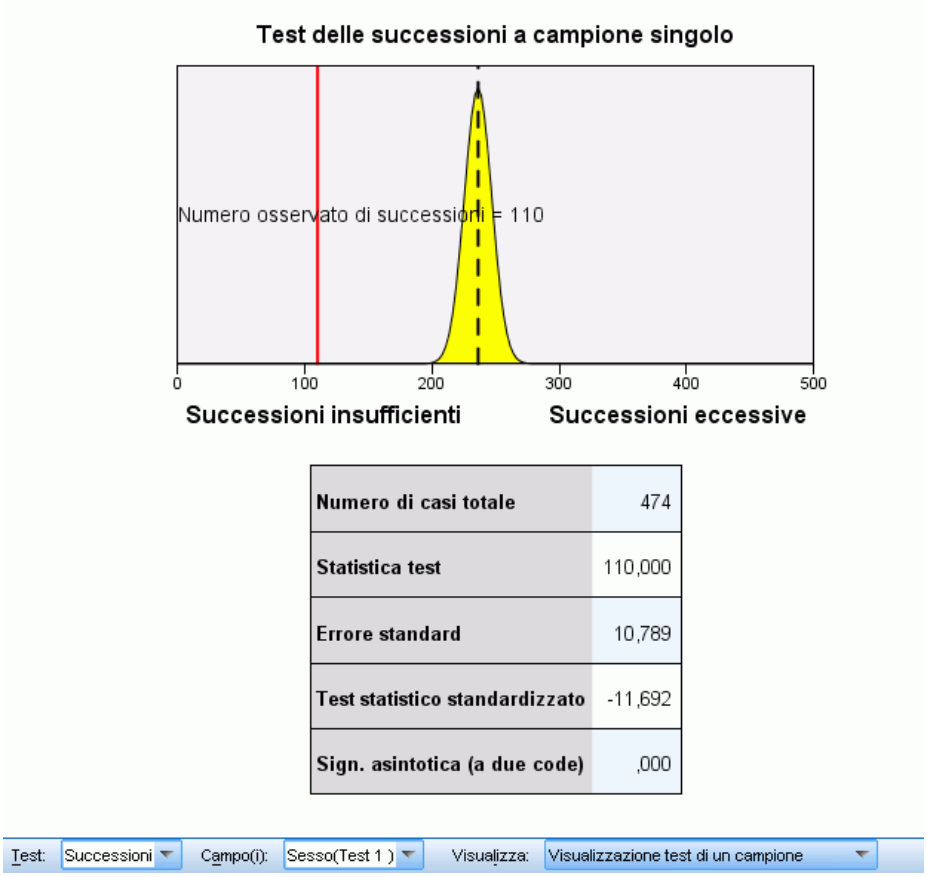


La visualizzazione Test dei segni per ranghi di Wilcoxon mostra un istogramma e una tabella dei test.

- L'istogramma comprende delle righe verticali che mostrano la mediana ipotetica e osservata.
- La tabella mostra i dettagli del test.

Test delle successioni

Figura 27-29
Visualizzazione test di un campione, Test delle successioni

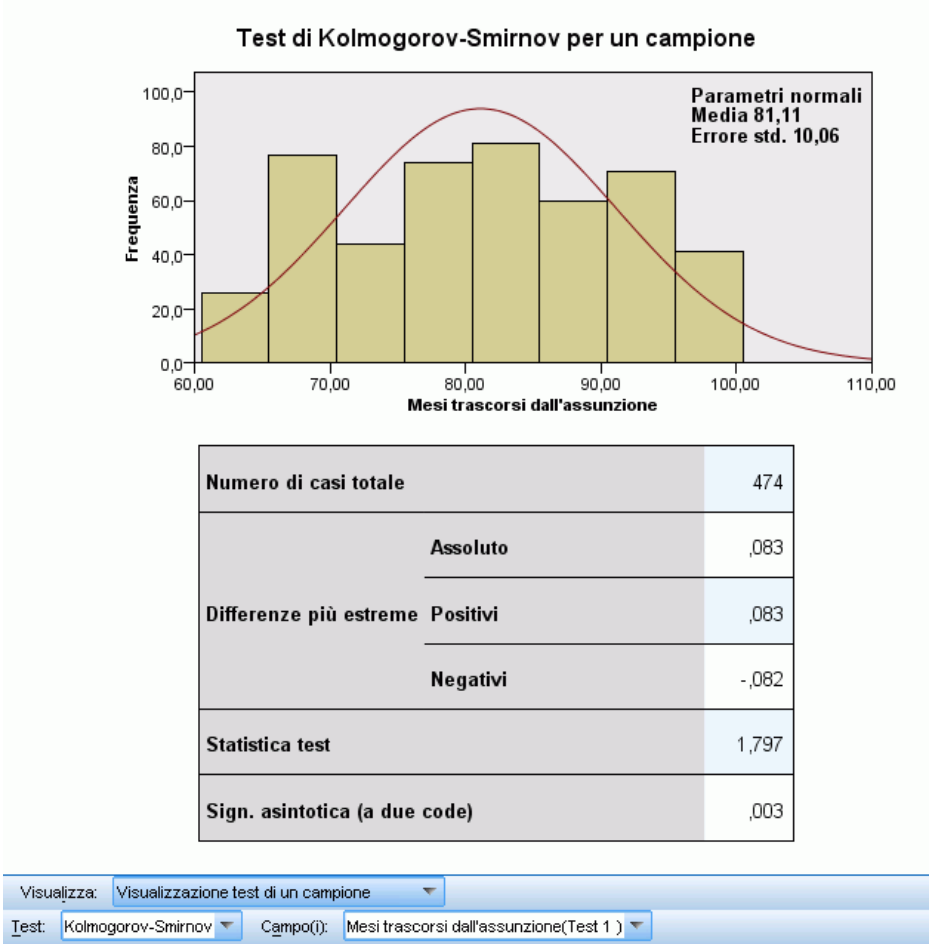


La visualizzazione Test delle successioni mostra un grafico e una tabella dei test.

- Il grafico mostra una distribuzione normale con il numero osservato di successioni contrassegnato da una linea verticale. Si noti che, quando si esegue il test esatto, il test non è basato sulla distribuzione normale.
- La tabella mostra i dettagli del test.

Test di Kolmogorov-Smirnov

Figura 27-30
Visualizzazione test di un campione, Test di Kolmogorov-Smirnov



La visualizzazione Test di Kolmogorov-Smirnov mostra un istogramma e una tabella dei test.

- L'istogramma comprende una sovrapposizione della funzione di densità di probabilità per la distribuzione ipotizzata uniforme, normale, di Poisson o esponenziale. Si noti che il test si basa su distribuzioni cumulative e le Differenze più estreme riportate nella tabella devono essere interpretate in relazione alle distribuzioni cumulative.
- La tabella mostra i dettagli del test.

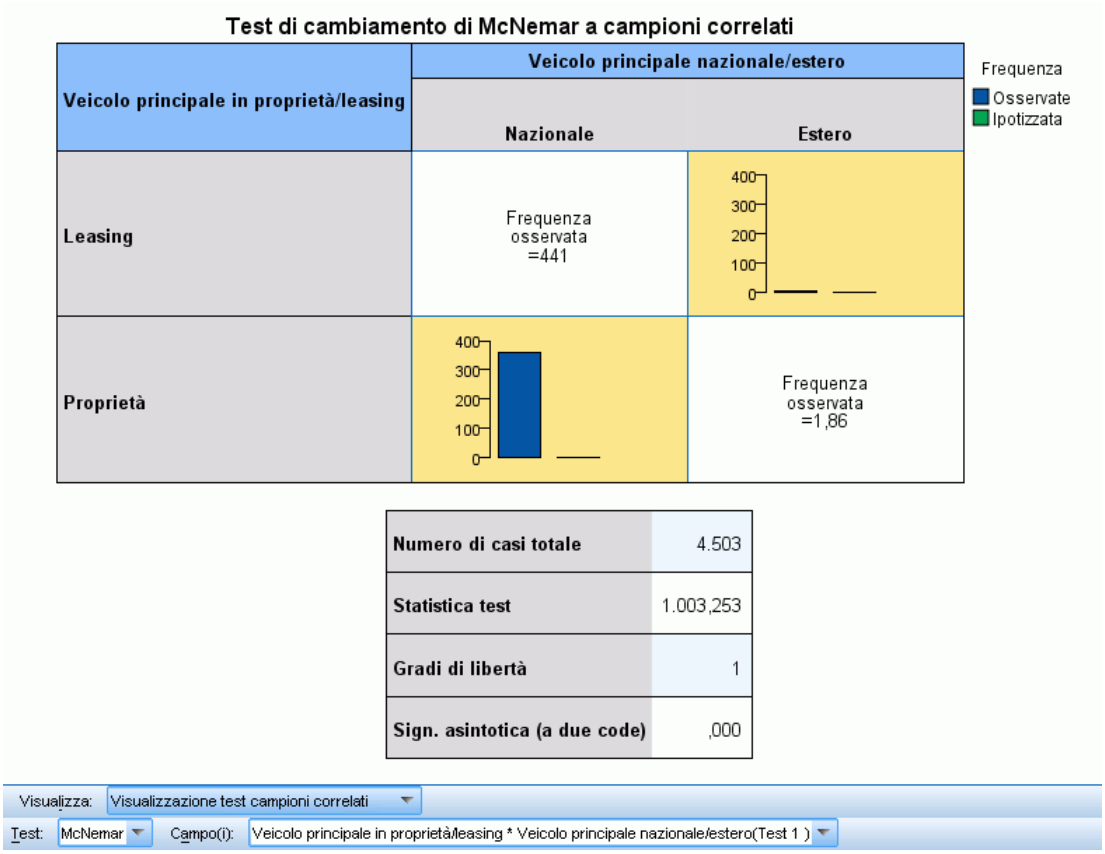
Test campioni correlati

La Visualizzazione test di un campione mostra i dettagli relativi a tutti i test non parametrici a campione singolo richiesti. Le informazioni visualizzate dipendono dal test selezionato.

- L'elenco a discesa Test consente di selezionare il tipo di test a campione singolo desiderato.
- L'elenco a discesa Campo(i) consente di selezionare un campo sottoposto al test selezionato nell'elenco a discesa Test.

Test di McNemar

Figura 27-31
Visualizzazione Test campioni correlati, Test di McNemar

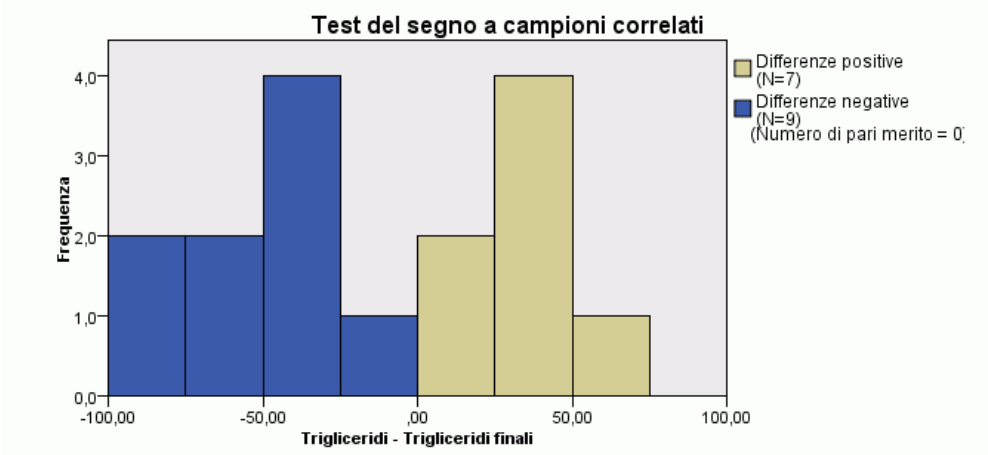


La visualizzazione Test di McNemar mostra un grafico a barre raggruppato e una tabella dei test.

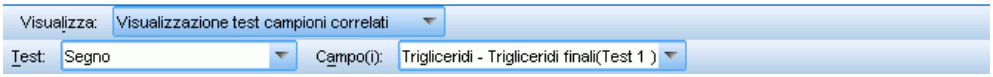
- Il grafico a barre raggruppato mostra le frequenze osservate e ipotizzate per le celle esterne alla diagonale della tabella 2×2 definita dai campi di test.
- La tabella mostra i dettagli del test.

Test del segno

Figura 27-32
Visualizzazione test campioni correlati, Test del segno



Numero di casi totale	16
Statistica test	7,000
Errore standard	2,000
Test statistico standardizzato	-,250
Sign. asintotica (a due code)	,803
Sign. esatta (a due code)	,804

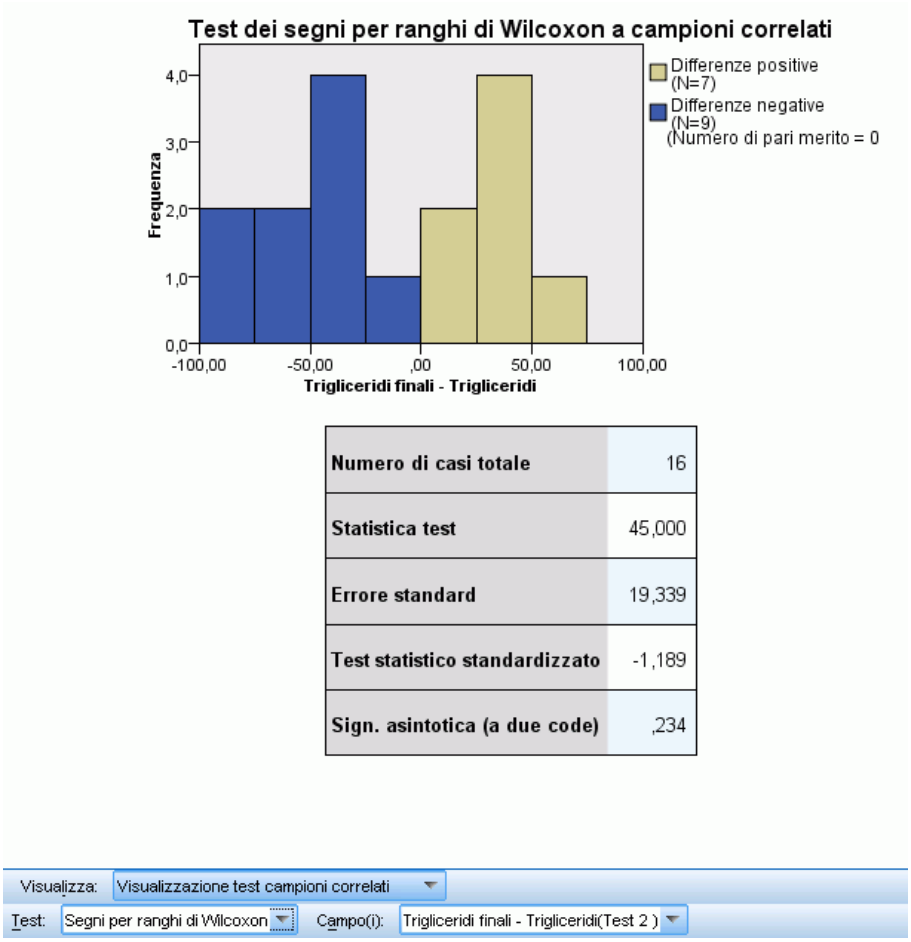


La visualizzazione Test del segno mostra un istogramma sovrapposto e una tabella dei test.

- L'istogramma sovrapposto mostra le differenze tra i campi utilizzando il segno della differenza come campo di sovrapposizione.
- La tabella mostra i dettagli del test.

Test dei segni per ranghi di Wilcoxon

Figura 27-33
Visualizzazione test campioni correlati, Test dei segni per ranghi di Wilcoxon

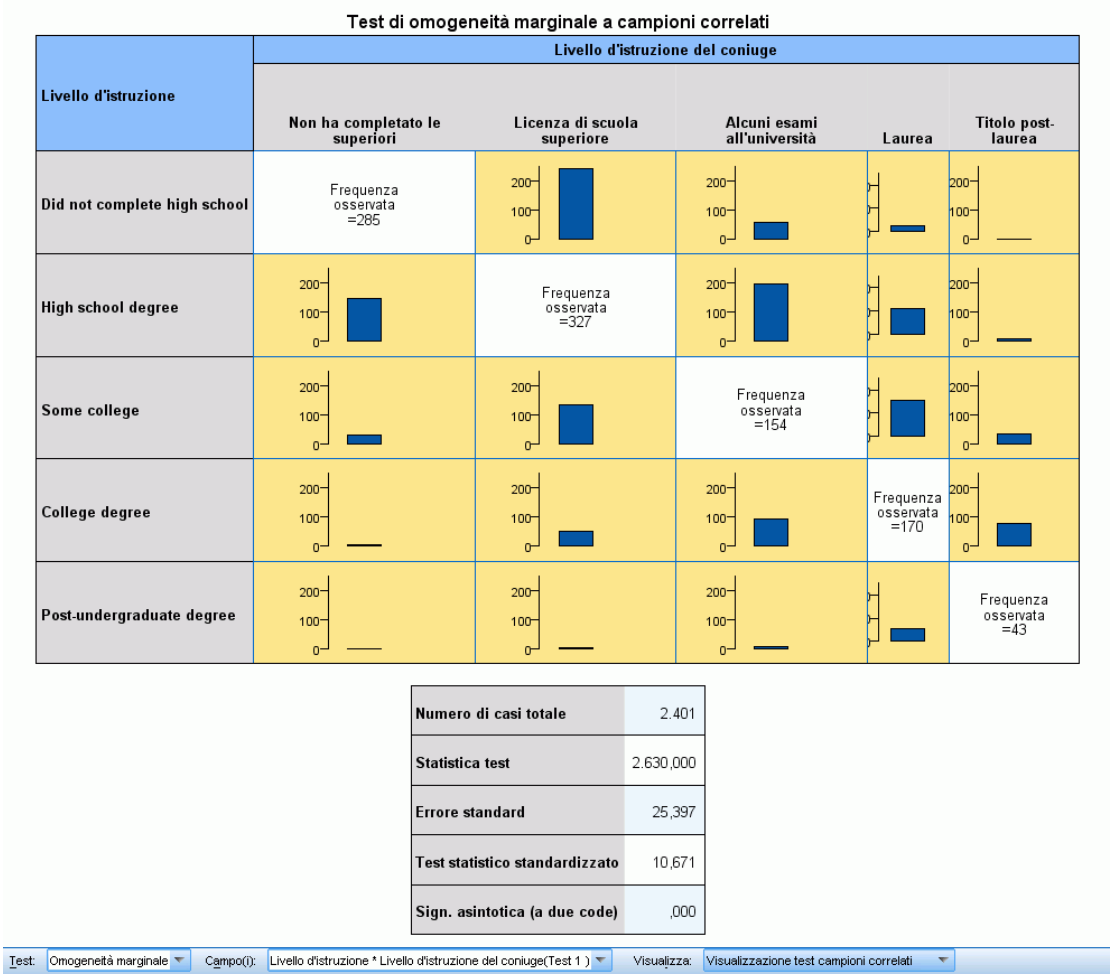


La visualizzazione Test dei segni per ranghi di Wilcoxon mostra un istogramma sovrapposto e una tabella dei test.

- L'istogramma sovrapposto mostra le differenze tra i campi utilizzando il segno della differenza come campo di sovrapposizione.
- La tabella mostra i dettagli del test.

Test di omogeneità marginale

Figura 27-34
Visualizzazione test campioni correlati, Test di omogeneità marginale

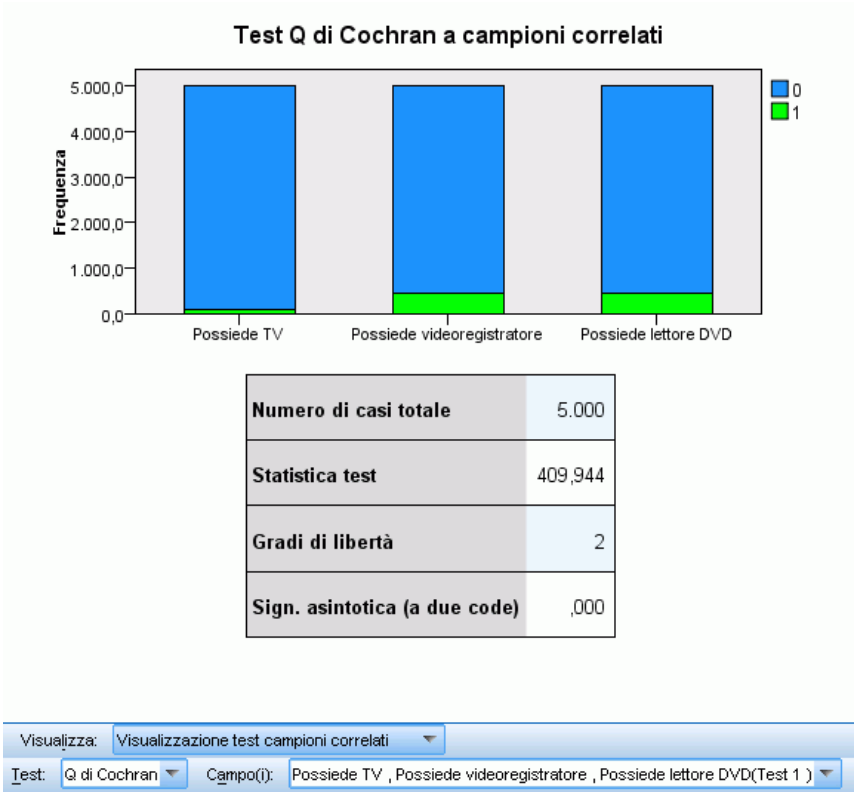


La visualizzazione Test di omogeneità marginale mostra un grafico a barre raggruppato e una tabella dei test.

- Il grafico a barre raggruppato mostra le frequenze osservate per le celle esterne alla diagonale della tabella definita dai campi di test.
- La tabella mostra i dettagli del test.

Test Q di Cochran

Figura 27-35
Visualizzazione test campioni correlati, Test Q di Cochran



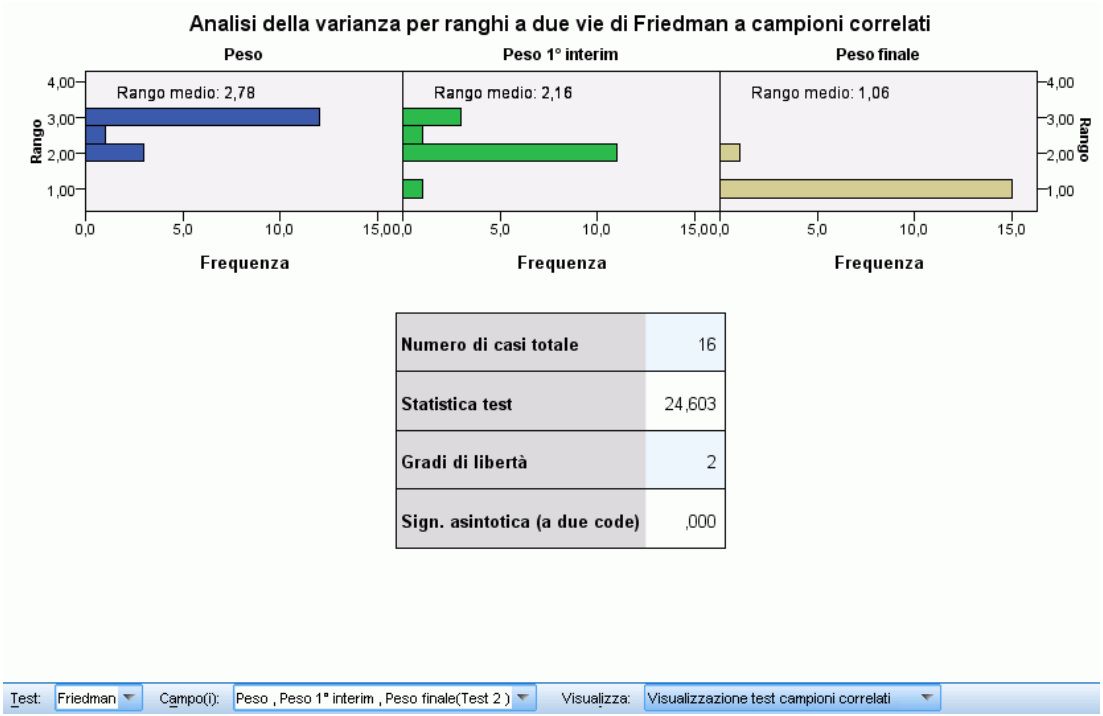
Numero di casi totale	5.000
Statistica test	409,944
Gradi di libertà	2
Sign. asintotica (a due code)	,000

La visualizzazione Test Q di Cochran mostra un grafico a barre sovrapposto e una tabella dei test.

- Il grafico a barre sovrapposto mostra le frequenze osservate per le categorie “esito positivo” ed “esito negativo” dei campi di test, con gli “esiti negativi” sovrapposti agli “esiti positivi”. Se si passa il mouse sopra una barra viene visualizzata una descrizione con le percentuali della categoria.
- La tabella mostra i dettagli del test.

Analisi della varianza per ranghi a due vie di Friedman

Figura 27-36
Visualizzazione test campioni correlati, Analisi della varianza per ranghi a due vie di Friedman

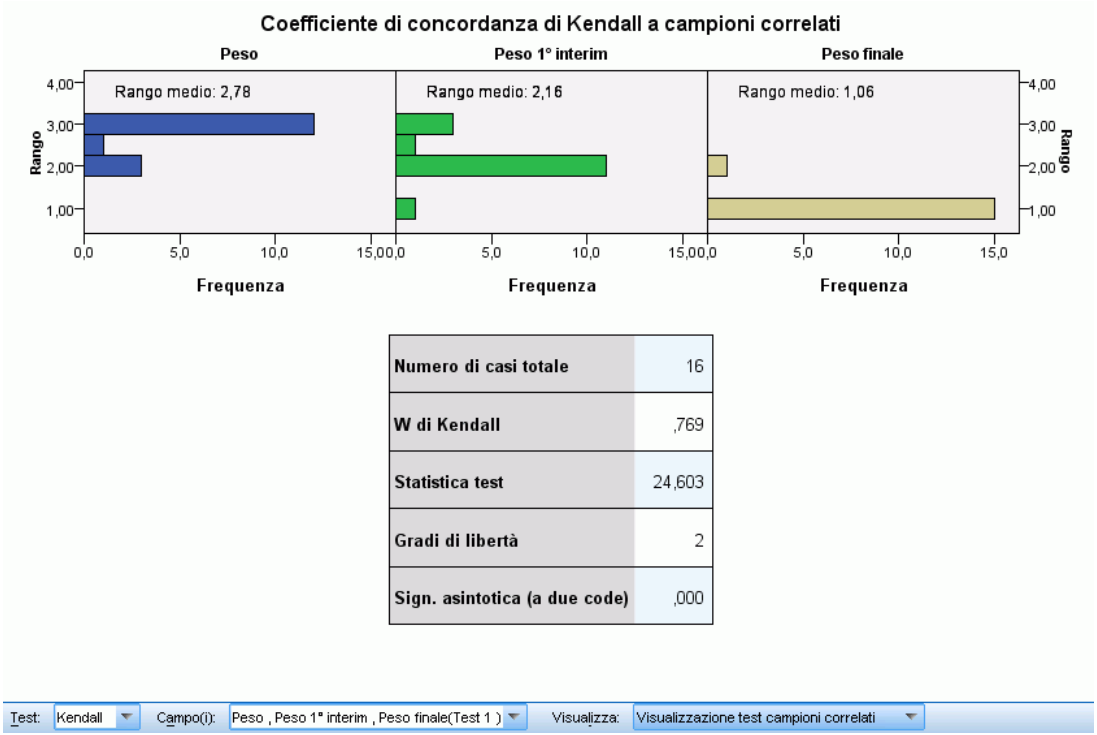


La visualizzazione Analisi della varianza per ranghi a due vie di Friedman mostra degli istogrammi a riquadri e una tabella dei test.

- Gli istogrammi mostrano la distribuzione osservata dei ranghi, suddivisa in riquadri in base ai campi di test.
- La tabella mostra i dettagli del test.

Coefficiente di concordanza di Kendall

Figura 27-37
Visualizzazione test campioni correlati, Coefficiente di concordanza di Kendall



La visualizzazione Coefficiente di concordanza di Kendall mostra degli istogrammi a riquadri e una tabella dei test.

- Gli istogrammi mostrano la distribuzione osservata dei ranghi, suddivisa in riquadri in base ai campi di test.
- La tabella mostra i dettagli del test.

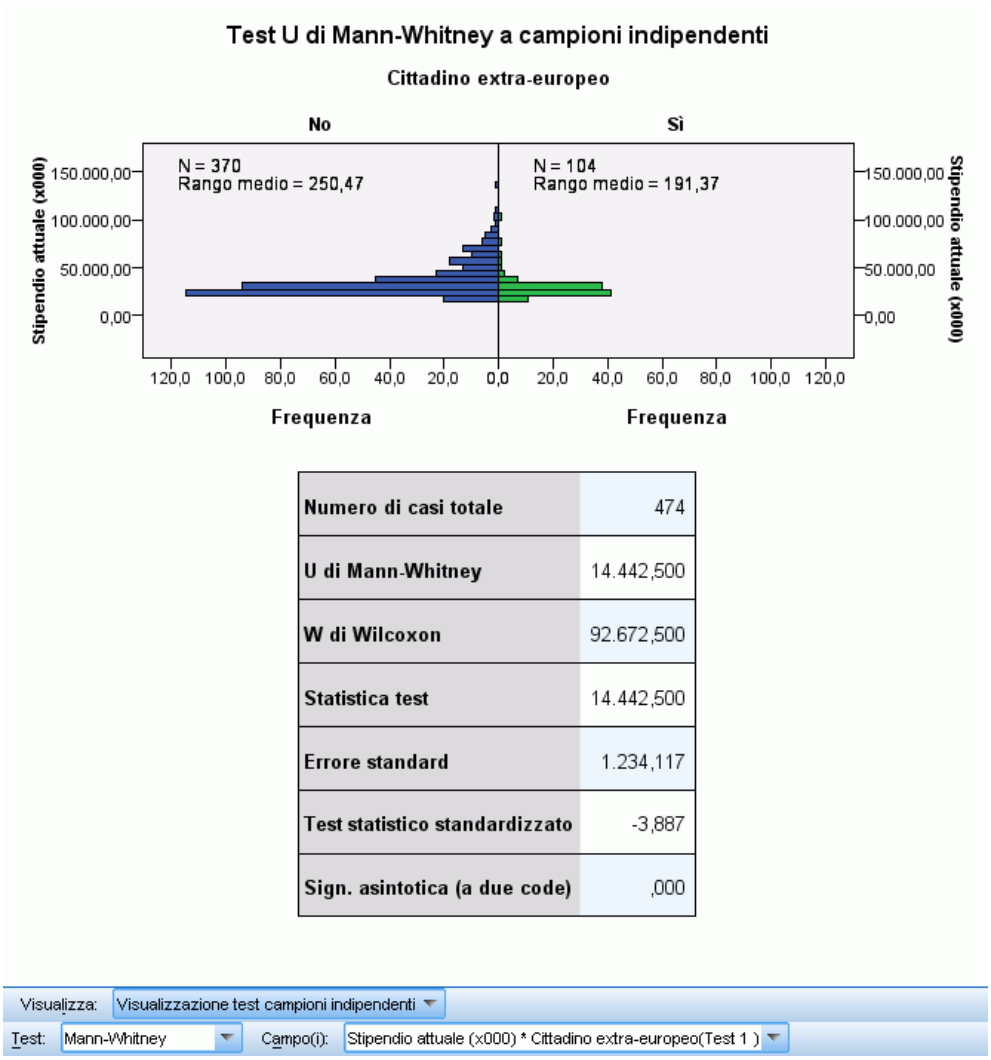
Test campioni indipendenti

La Visualizzazione test campioni indipendenti mostra i dettagli relativi a tutti i test non parametrici per campioni indipendenti richiesti. Le informazioni visualizzate dipendono dal test selezionato.

- L'elenco a discesa Test consente di selezionare il tipo di test per campioni indipendenti desiderato.
- L'elenco a discesa Campo(i) consente di selezionare una combinazione di campi di test e di raggruppamento sottoposta al test selezionato nell'elenco a discesa Test.

Test di Mann-Whitney

Figura 27-38
Visualizzazione test campioni indipendenti, Test di Mann-Whitney



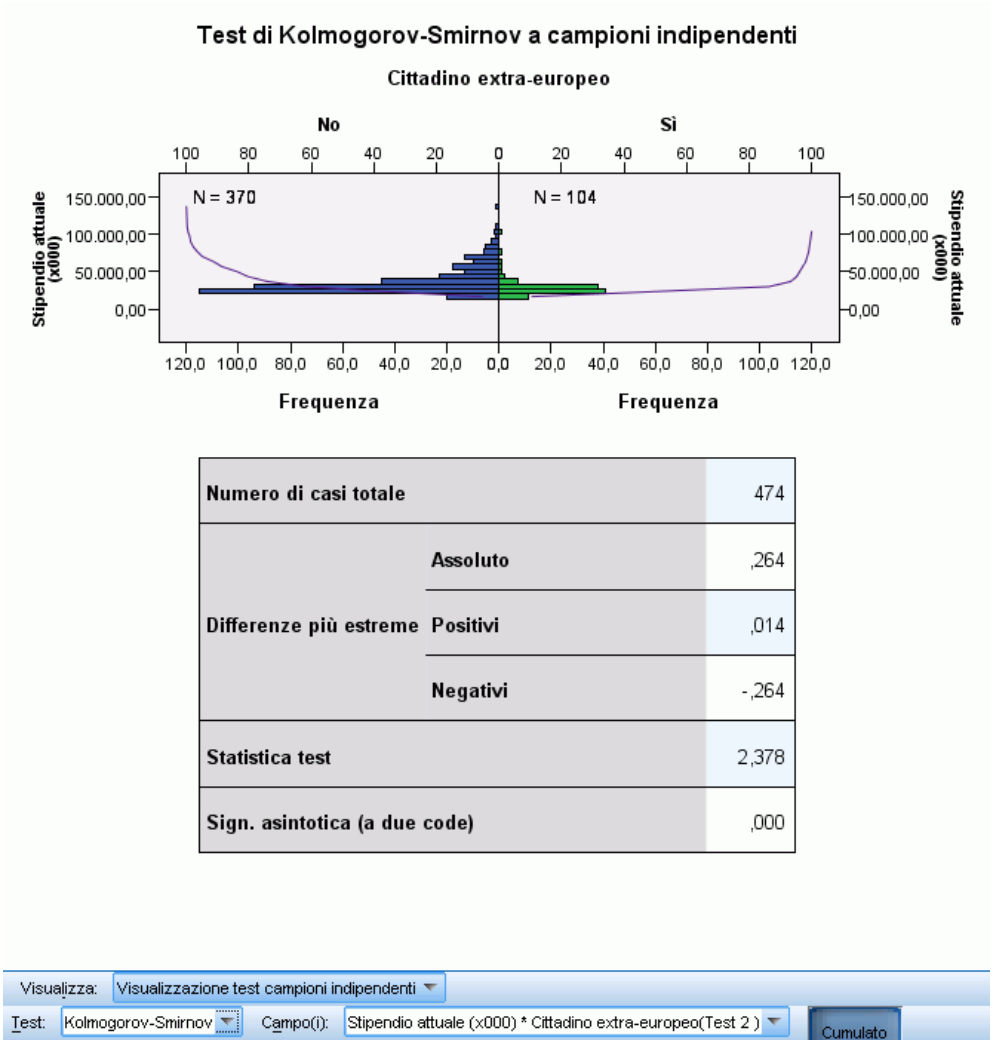
Numero di casi totale	474
U di Mann-Whitney	14.442,500
W di Wilcoxon	92.672,500
Statistica test	14.442,500
Errore standard	1.234,117
Test statistico standardizzato	-3,887
Sign. asintotica (a due code)	,000

La visualizzazione Test di Mann-Whitney mostra un grafico a piramide della popolazione e una tabella dei test.

- Il grafico a piramide della popolazione mostra degli istogrammi affiancati in modalità retroversa in base alle categorie del campo di raggruppamento, notando il numero di record di ogni gruppo e il rango medio del gruppo.
- La tabella mostra i dettagli del test.

Test di Kolmogorov-Smirnov

Figura 27-39
Visualizzazione test campioni indipendenti, Test di Kolmogorov-Smirnov



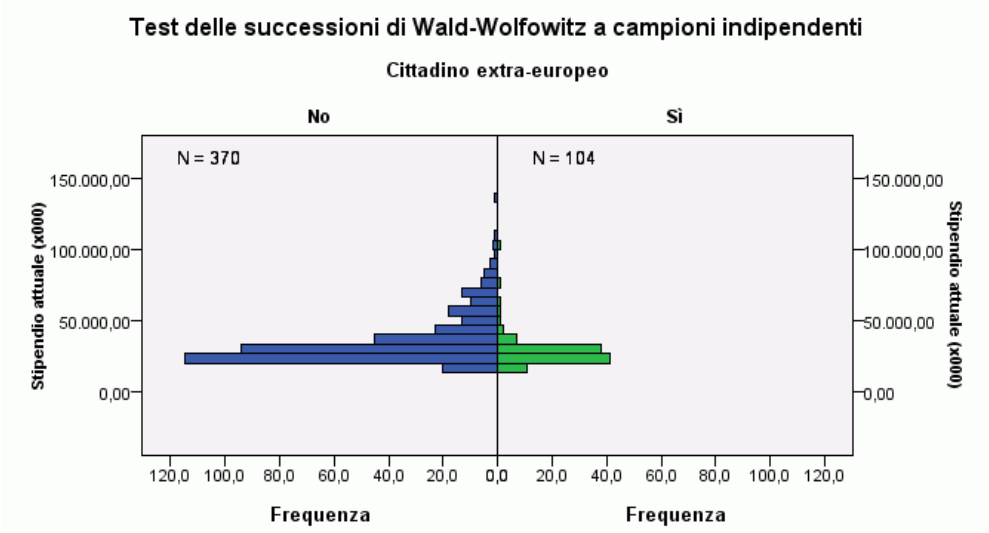
Numero di casi totale		474
Differenze più estreme	**Assoluto**	,264
	Positivi	,014
	Negativi	-,264
Statistica test		2,378
Sign. asintotica (a due code)		,000

La visualizzazione Test di Kolmogorov-Smirnov mostra un grafico a piramide della popolazione e una tabella dei test.

- Il grafico a piramide della popolazione mostra degli istogrammi affiancati in modalità retroversa in base alle categorie del campo di raggruppamento, notando il numero di record di ogni gruppo. Le linee della distribuzione cumulativa osservata si possono visualizzare o nascondere facendo clic sul pulsante Cumulato.
- La tabella mostra i dettagli del test.

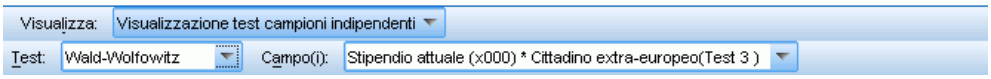
Test delle successioni di Wald-Wolfowitz

Figura 27-40
Visualizzazione test campioni indipendenti, Test delle successioni di Wald-Wolfowitz



Numero di casi totale		474
Minimo possibile	**Statistica test**[1]	97,000
	Errore standard	7,442
	Test statistico standardizzato	-8,917
	Sign. asintotica (a due code)	,000
Massimo possibile	**Statistica test**[1]	199,000
	Errore standard	7,442
	Test statistico standardizzato	4,788
	Sign. asintotica (a due code)	1,000

[1]The test statistic is the number of runs.
1. There are 55 inter-group ties involving 228 records.

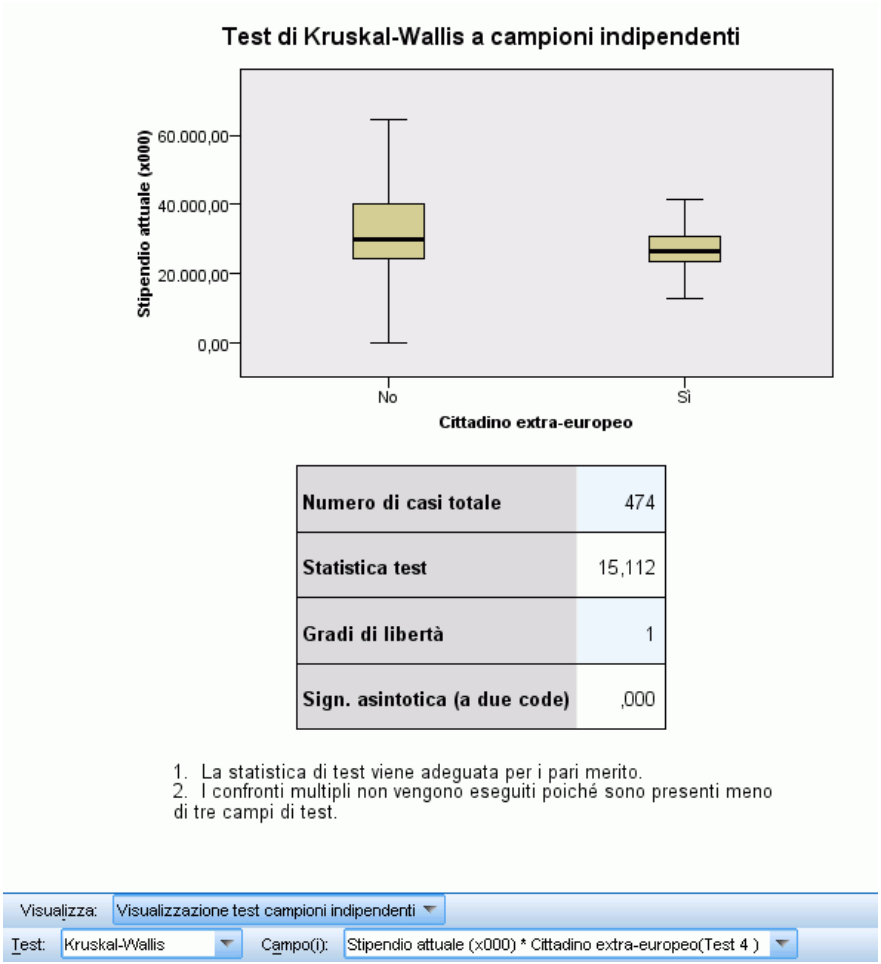


La visualizzazione Test delle successioni di Wald-Wolfowitz mostra un grafico a barre sovrapposto e una tabella dei test.

- Il grafico a piramide della popolazione mostra degli istogrammi affiancati in modalità retroversa in base alle categorie del campo di raggruppamento, notando il numero di record di ogni gruppo.
- La tabella mostra i dettagli del test.

Test di Kruskal-Wallis

Figura 27-41
Visualizzazione test campioni indipendenti, Test di Kruskal-Wallis



Numero di casi totale	474
Statistica test	15,112
Gradi di libertà	1
Sign. asintotica (a due code)	,000

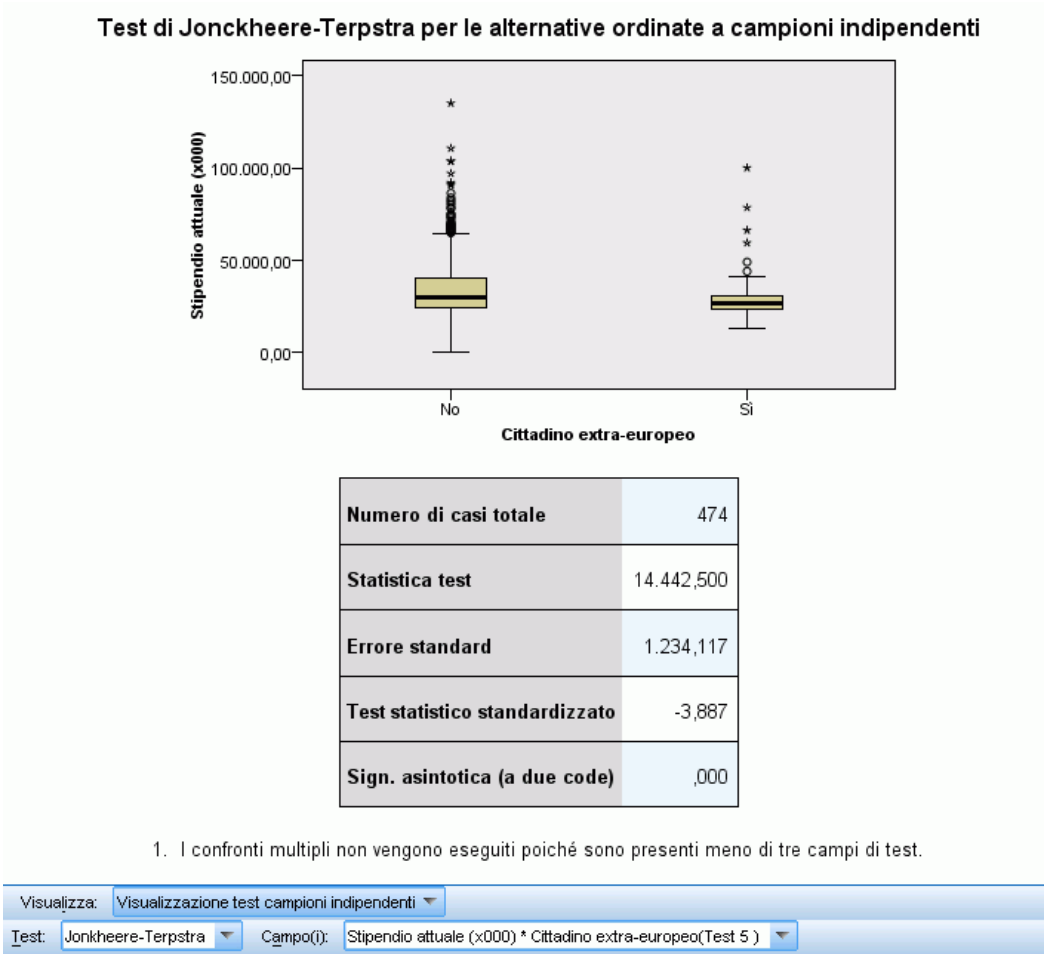
1. La statistica di test viene adeguata per i pari merito.
2. I confronti multipli non vengono eseguiti poiché sono presenti meno di tre campi di test.

La visualizzazione Test Kruskal-Wallis mostra dei grafici a scatole e una tabella dei test.

- Per ogni categoria del campo di raggruppamento vengono visualizzati grafici a scatole separati. Se si passa il mouse sopra una scatola viene visualizzata una descrizione con il rango medio.
- La tabella mostra i dettagli del test.

Test di Jonckheere-Terpstra

Figura 27-42
Visualizzazione test campioni indipendenti, Test di Jonckheere-Terpstra



Numero di casi totale	474
Statistica test	14.442,500
Errore standard	1.234,117
Test statistico standardizzato	-3,887
Sign. asintotica (a due code)	,000

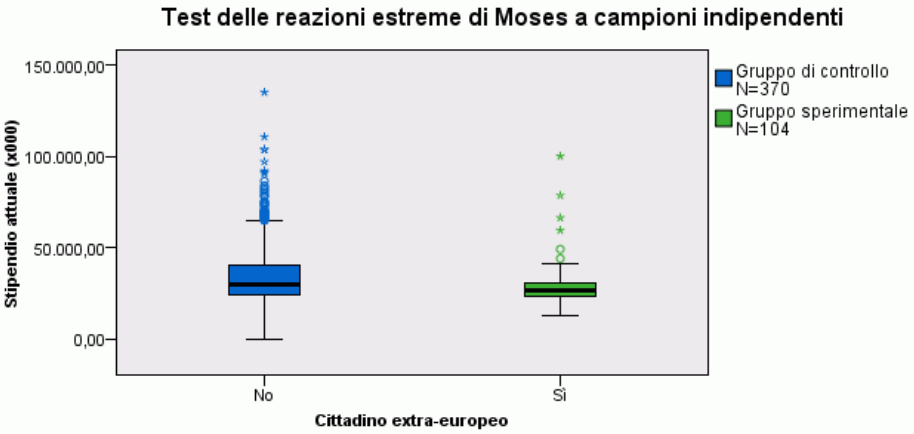
1. I confronti multipli non vengono eseguiti poiché sono presenti meno di tre campi di test.

La visualizzazione Test di Jonckheere-Terpstra mostra dei grafici a scatole e una tabella dei test.

- Per ogni categoria del campo di raggruppamento vengono visualizzati grafici a scatole separati.
- La tabella mostra i dettagli del test.

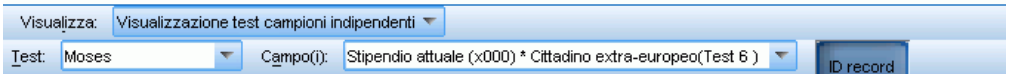
Test delle reazioni estreme di Moses

Figura 27-43
Visualizzazione test campioni indipendenti, Test delle reazioni estreme di Moses



Numero di casi totale		474
Gruppo di controllo osservato	**Statistica test[1]**	474,000
	Sign. esatta (a una coda)	1,000
Gruppo di controllo tagliato	**Statistica test[1]**	423,000
	Sign. esatta (a una coda)	,341
Valori anomali tagliati da ogni estremità		19,000

[1]The test statistic is the span.

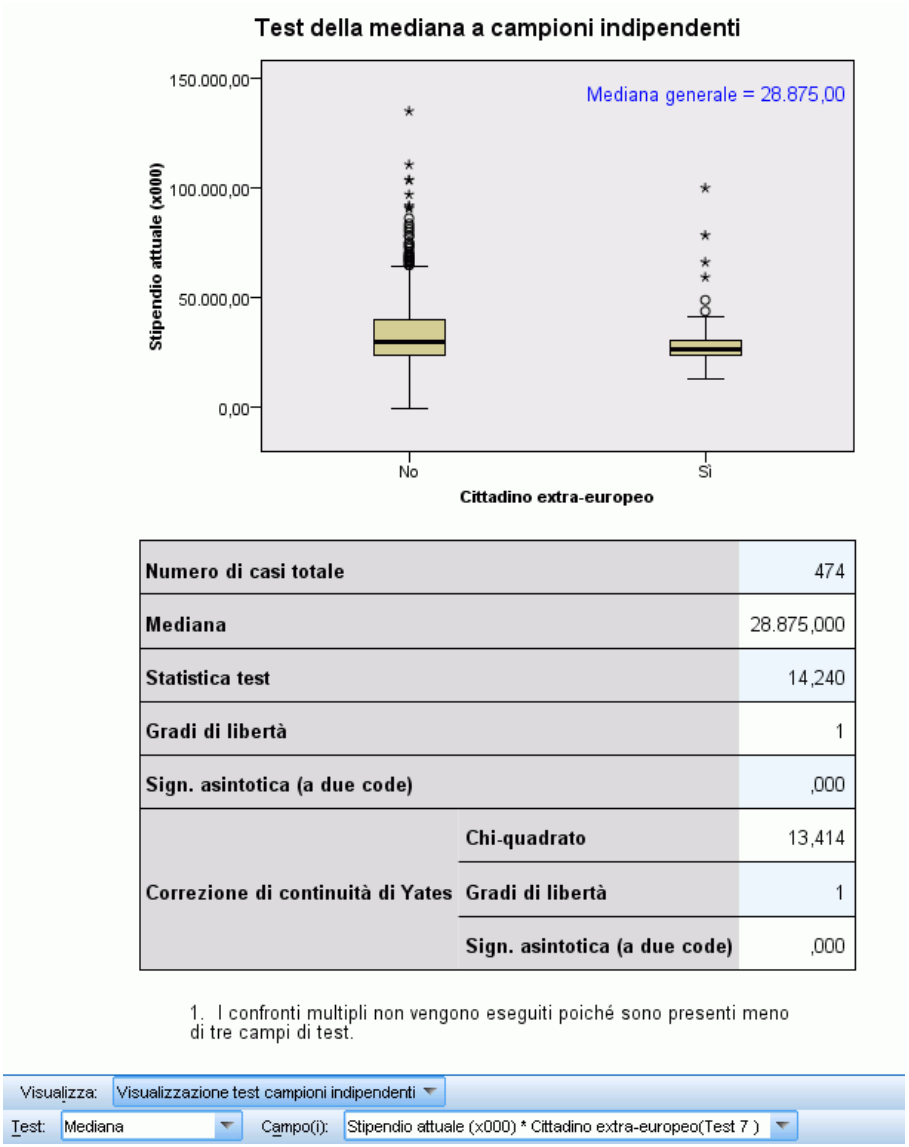


La visualizzazione Test delle reazioni estreme di Moses mostra dei grafici a scatole e una tabella dei test.

- Per ogni categoria del campo di raggruppamento vengono visualizzati grafici a scatole separati. Le etichette dei punti si possono visualizzare o nascondere facendo clic sul pulsante ID record.
- La tabella mostra i dettagli del test.

Test della mediana

Figura 27-44
Visualizzazione test campioni indipendenti, Test della mediana



Numero di casi totale		474
Mediana		28.875,000
Statistica test		14,240
Gradi di libertà		1
Sign. asintotica (a due code)		,000
Correzione di continuità di Yates	**Chi-quadrato**	13,414
	Gradi di libertà	1
	Sign. asintotica (a due code)	,000

1. I confronti multipli non vengono eseguiti poiché sono presenti meno di tre campi di test.

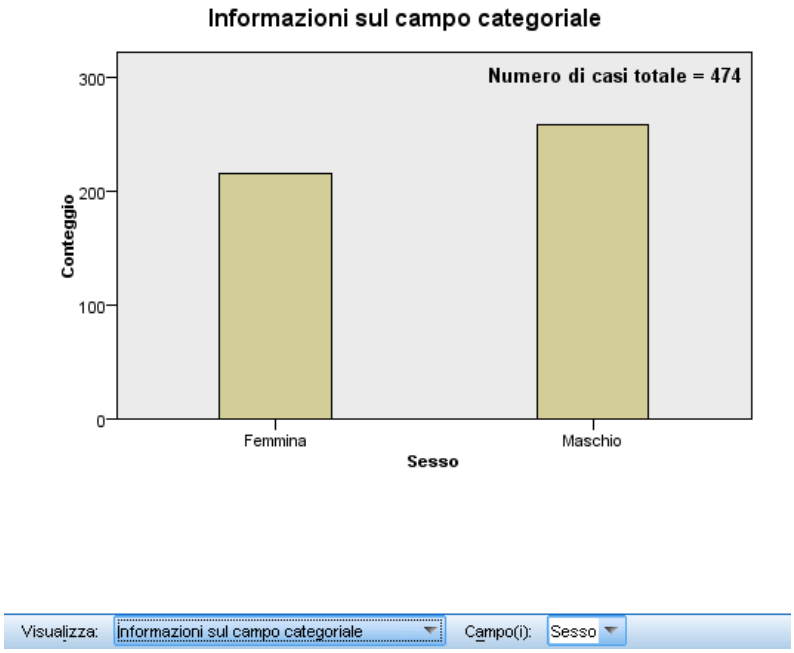
Visualizza: Visualizzazione test campioni indipendenti
Test: Mediana Campo(i): Stipendio attuale (x000) * Cittadino extra-europeo(Test 7)

La visualizzazione Test della mediana mostra dei grafici a scatole e una tabella dei test.

- Per ogni categoria del campo di raggruppamento vengono visualizzati grafici a scatole separati.
- La tabella mostra i dettagli del test.

Informazioni sul campo categoriale

Figura 27-45
Informazioni sul campo categoriale

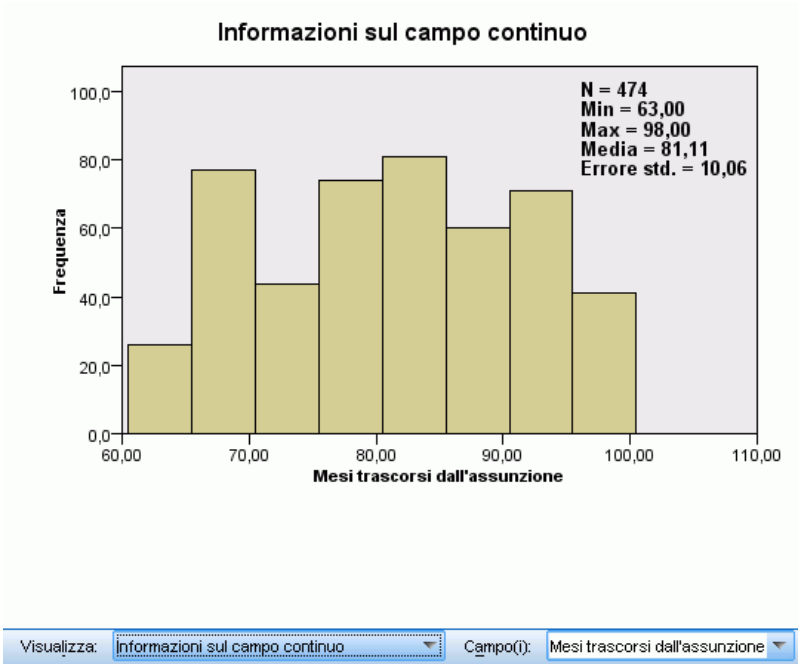


La visualizzazione Informazioni sul campo categoriale mostra un grafico a barre per il campo categoriale selezionato nell'elenco a discesa Campo(i). L'elenco dei campi disponibili è limitato ai campi categoriali utilizzati nel test selezionato nella Visualizzazione riepilogo ipotesi.

- Se si passa il mouse sopra una barra viene visualizzata una descrizione con le percentuali della categoria.

Informazioni sul campo continuo

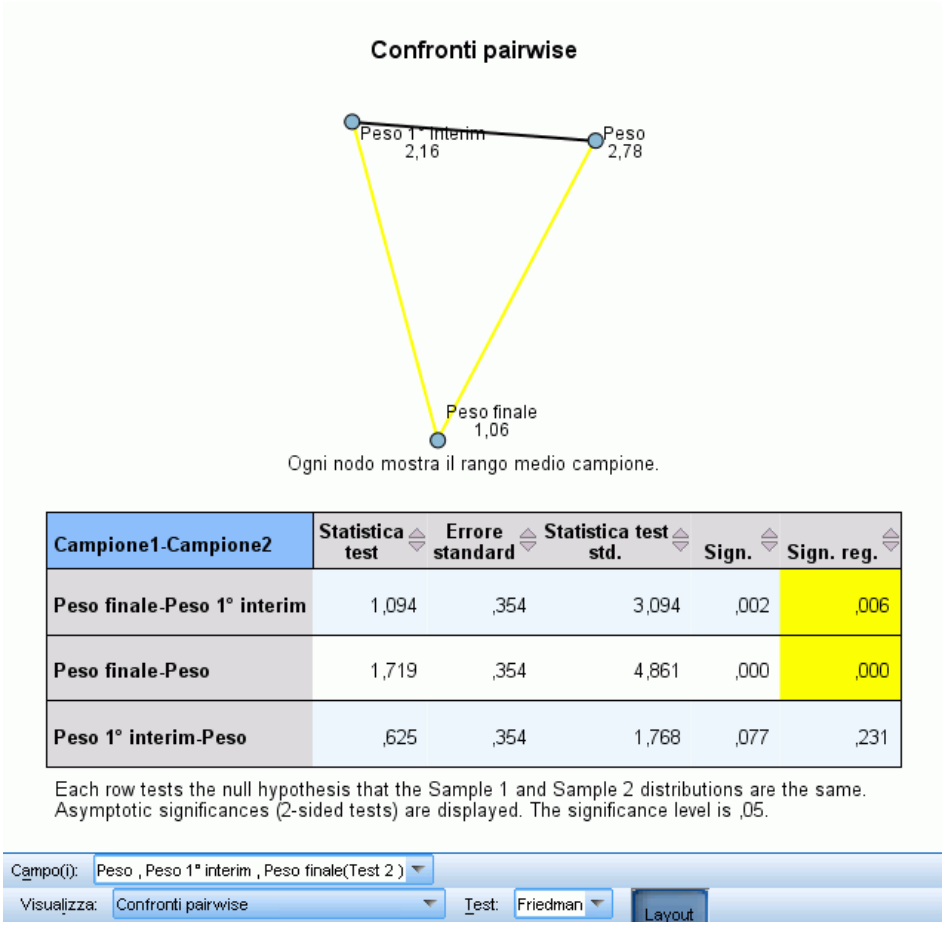
Figura 27-46
Informazioni sul campo continuo



La visualizzazione Informazioni sul campo continuo mostra un istogramma per il campo continuo selezionato nell'elenco a discesa Campo(i). L'elenco dei campi disponibili è limitato ai campi continui utilizzati nel test selezionato nella Visualizzazione riepilogo ipotesi.

Confronti pairwise

Figura 27-47
Confronti pairwise



Campione1-Campione2	Statistica test	Errore standard	Statistica test std.	Sign.	Sign. reg.
Peso finale-Peso 1° interim	1,094	,354	3,094	,002	,006
Peso finale-Peso	1,719	,354	4,861	,000	,000
Peso 1° interim-Peso	,625	,354	1,768	,077	,231

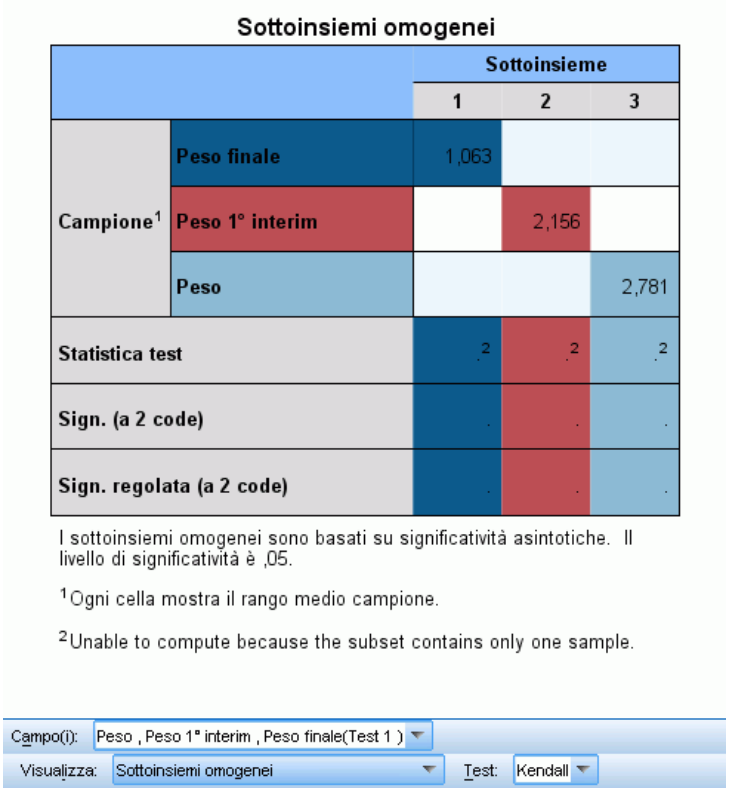
Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is ,05.

La visualizzazione Confronti pairwise mostra un grafico di rete delle distanze e una tabella dei confronti generati da test non parametrici *K*-S quando vengono richiesti più confronti pairwise.

- Il grafico di rete delle distanze è una rappresentazione grafica della tabella dei confronti in cui le distanze fra i nodi della rete corrispondono a differenze tra i campioni. Le linee gialle corrispondono alle differenze statisticamente significative; le linee nere corrispondono alle differenze non significative. Se si passa il mouse sopra una linea della rete viene visualizzata una descrizione con la significatività corretta della differenza tra i nodi collegati dalla linea.
- La tabella dei confronti mostra i risultati numerici di tutti i confronti pairwise. Ogni riga corrisponde a un confronto pairwise diverso. Fare clic sull'intestazione di una colonna per ordinare le righe in base ai valori di quella colonna.

Sottoinsiemi omogenei

Figura 27-48
Sottoinsiemi omogenei



Sottoinsiemi omogenei

		Sottoinsieme		
		1	2	3
Campione[1]	Peso finale	1,063		
	Peso 1° interim		2,156	
	Peso			2,781
Statistica test		.[2]	.[2]	.[2]
Sign. (a 2 code)		.	.	.
Sign. regolata (a 2 code)		.	.	.

I sottoinsiemi omogenei sono basati su significatività asintotiche. Il livello di significatività è ,05.

[1] Ogni cella mostra il rango medio campione.

[2] Unable to compute because the subset contains only one sample.

La visualizzazione Sottoinsiemi omogenei mostra una tabella dei confronti generata da test non parametrici *K*-S quando vengono richiesti più confronti stepwise stepdown.

- Ogni riga del gruppo Campione corrisponde a un campione correlato separato (rappresentato nei dati da campi distinti). I campioni che non presentano differenze statisticamente significative sono raggruppati in sottoinsiemi dello stesso colore, con una colonna separata per ogni sottoinsieme identificato. Quando tutti i campioni presentano differenze statisticamente significative, viene creato un sottoinsieme separato per ogni campione. Quando nessuno dei campioni presenta differenze statisticamente significative, il sottoinsieme è unico.
- Per ogni sottoinsieme contenente più di un campione vengono calcolati la statistica del test, il valore di significatività e di significatività corretta.

Opzioni aggiuntive del comando NPTESTS

Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare test a campione singolo, a campioni indipendenti e a campioni correlati in un'unica esecuzione della procedura.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Finestre legacy

Esistono anche numerose finestre di dialogo "legacy" che eseguono test non parametrici. Queste finestre di dialogo supportano la funzionalità offerta dall'opzione Test esatti.

Test Chi-quadrato. Consente di analizzare una variabile in categorie e di ottenere una statistica del chi-quadrato in base alle differenze tra frequenze osservate e attese.

Test binomiale. Consente di confrontare la frequenza osservata in ciascuna categoria di una variabile dicotomica con le frequenze attese dalla distribuzione binomiale.

Test delle successioni. Consente di verificare se l'ordine di occorrenza di due valori di una variabile è casuale.

Test di Kolmogorov-Smirnov per un campione. Consente di confrontare la funzione di distribuzione cumulata osservata per una variabile con la distribuzione teorica specificata, che può essere normale, uniforme, esponenziale o di Poisson.

Test per due campioni indipendenti. Consente di confrontare due gruppi di casi in base a una sola variabile. Sono disponibili il test *U* di Mann-Whitney, il test di Kolmogorov-Smirnov per due campioni, il test delle reazioni estreme di Moses e il test delle successioni di Wald-Wolfowitz.

Test per due campioni dipendenti. Consente di confrontare le distribuzioni di due variabili. Sono disponibili il test dei segni per ranghi di Wilcoxon, il test del segno e il test di McNemar.

Test per diversi campioni indipendenti. Consente di confrontare due o più gruppi di casi in base alla stessa variabile. Sono disponibili il test di Kruskal-Wallis, il test della mediana e il test di Jonckheere-Terpstra.

Test per diversi campioni correlati. Consente di confrontare le distribuzioni di due o più variabili. Sono disponibili il test di Friedman, il test *W* di Kendall e il test *Q* di Cochran.

Per tutti i test precedentemente citati sono disponibili quartili e media, deviazione standard, valore minimo e massimo e numero di casi non mancanti.

Test Chi-quadrato

La procedura Test chi-quadrato permette di analizzare una variabile in categorie e di calcolare una statistica chi-quadrato. Questo test sulla bontà di adattamento permette di confrontare le frequenze osservate e attese in ciascuna categoria per verificare se tutte le categorie includono la stessa proporzione di valori o se includono una proporzione di valori specificati dall'utente.

Esempi. Il test chi-quadrato può essere utilizzato per determinare se in un sacchetto di gelatine di frutta è presente la stessa proporzione di blu, marrone, arancio, rosso e giallo. È inoltre possibile determinare se il sacchetto di gelatine contiene il 5% di blu, il 30% di marrone, il 10% di verde, il 20% di arancio, il 15% di rosso e il 15% di giallo.

Statistiche. Media, deviazione standard, valore minimo e massimo e quartili. Il numero e la percentuale di casi mancanti e non mancanti, il numero di casi osservati e attesi per ciascuna categoria, i residui e la statistica chi-quadrato.

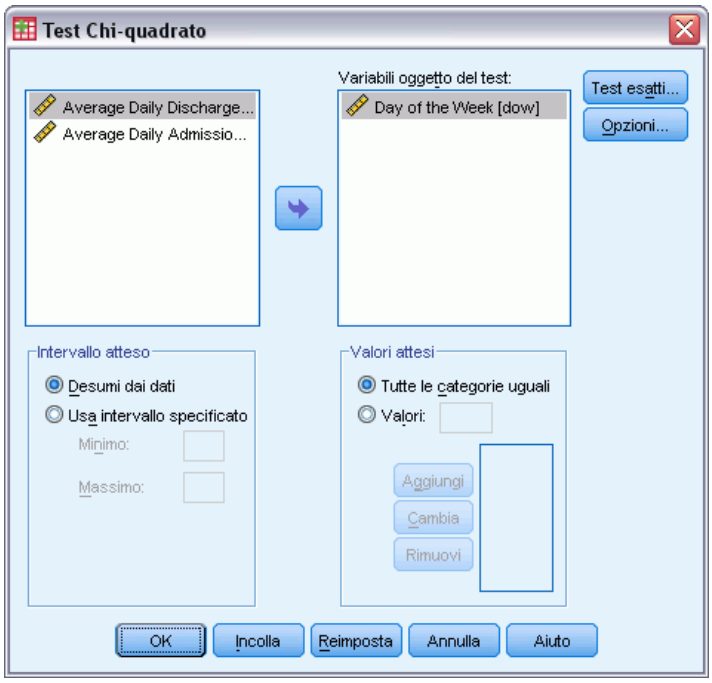
Dati. Utilizzare variabili categoriali numeriche ordinate o non ordinate (livelli di misurazione ordinali o nominali). Per convertire le variabili stringa in variabili numeriche, utilizzare il comando Ricodifica automatica del menu Trasforma.

Assunzioni. I test non parametrici non richiedono ipotesi relative alla forma della distribuzione sottostante. Si presume che i dati rappresentino un campione casuale. Le frequenze attese per ciascuna categoria devono essere come minimo pari a 1. Non più del 20% delle categorie possono avere frequenze attese inferiori a 5.

Per ottenere un test chi-quadrato

- Dai menu, scegliere:
  Analizza > Test non parametrici > Finestre legacy > Chi-Quadrato...

Figura 27-49
Finestra di dialogo Test chi-quadrato



- Selezionare una o più variabili per il test. Ogni variabile produce un test distinto.
- È possibile fare clic su Opzioni per ottenere statistiche descrittive, quartili e controllo delle modalità di elaborazione dei dati mancanti.

Intervallo e valori attesi del test chi-quadrato

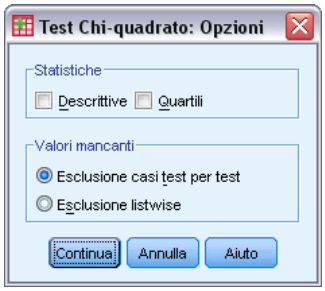
Intervallo atteso. Per impostazione predefinita, ogni singolo valore della variabile è definito come categoria. Per definire categorie all'interno di un intervallo specifico, selezionare Usaintervallo specificato e inserire valori interi per il limite inferiore e superiore. Le categorie verranno definite come valori inclusi nell'intervallo specificato. I casi con valori al di fuori del minimo e del massimo specificato saranno esclusi dal test. Se, ad esempio, si specifica 1 per il limite inferiore e

4 per il limite superiore, per il test chi-quadrato verranno utilizzati solo i valori interi compresi tra 1 e 4.

Valori attesi. Per impostazione predefinita, tutte le categorie hanno valori attesi uguali. Per le categorie sono previste proporzioni attese definite dall'utente. Selezionare Valori, specificare un valore maggiore di 0 per ogni categoria della variabile del test e quindi fare clic su Aggiungi. I valori vengono elencati in ordine di inserimento, L'ordine dei valori è importante in quanto corrisponde all'ordine crescente dei valori delle categorie della variabile oggetto del test. Il primo valore della lista corrisponde al valore di gruppo minore della variabile, mentre l'ultimo corrisponde al valore maggiore. Gli elementi della lista dei valori vengono sommati e quindi ciascun valore viene diviso per la somma risultante per calcolare la proporzione di casi attesi nella categoria corrispondente. Ad esempio, una lista valori formata da 3, 4, 5, 4 specifica le proporzioni attese 3/16, 4/16, 5/16 e 4/16.

Test chi-quadrato: Opzioni

Figura 27-50
Finestra di dialogo Test chi-quadrato: Opzioni



Statistiche. È possibile scegliere uno o entrambe le statistiche riassuntive.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile sono esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TESTS (test chi-quadrato)

Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare valori minimi e massimi diversi o frequenze attese diverse per variabili diverse (con il sottocomando `CHISQUARE`).
- Eseguire il test confrontando la stessa variabile con diverse frequenze attese o utilizzando diversi intervalli (con il sottocomando `EXPECTED`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Test binomiale

Grazie alla procedura del test binomiale è possibile confrontare le frequenze osservate delle due categorie di una variabile dicotomica con le frequenze previste in presenza di una distribuzione binomiale con il parametro di probabilità specificato. Per impostazione predefinita, il parametro di probabilità per entrambi i gruppi è 0,5. Per modificare le probabilità, è possibile inserire una proporzione di prova per il primo gruppo. La probabilità per il secondo gruppo sarà uguale a 1 meno la probabilità specificata per il primo gruppo.

Esempio. Quando si lancia in aria una moneta, la probabilità che esca testa è pari a 1/2. In base a questa ipotesi, la moneta viene lanciata in aria 40 volte e i risultati (testa o croce) vengono registrati. Dal test binomiale può risultare che per i 3/4 dei lanci della moneta è uscita testa e che il livello di significatività è molto basso (0,0027). Questi risultati indicano che la probabilità che venga testa molto spesso non è pari a 1/2; pertanto, la stima sul comportamento della moneta probabilmente risulta distorta.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili.

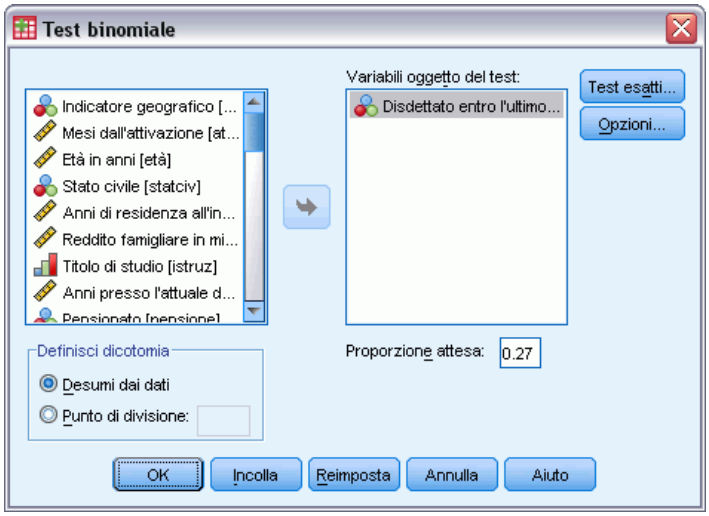
Dati. Le variabili incluse nel test devono essere numeriche e dicotomiche. Per convertire le variabili stringa in variabili numeriche, utilizzare il comando Ricodifica automatica del menu Trasforma. Una **variabile dicotomica** è una variabile che prevede solo due possibili valori: *Sì* o *No*, *Vero* o *Falso*, 0 o 1 e così via. Il primo valore rilevato nell'insieme di dati definisce il primo gruppo, mentre l'altro valore definisce il secondo gruppo. Se le variabili sono dicotomiche, è necessario specificare un punto di divisione. Utilizzando il punto di divisione è possibile assegnare al primo gruppo i casi con valori inferiori o uguali al punto di divisione e i rimanenti casi al secondo gruppo.

Assunzioni. I test non parametrici non richiedono ipotesi relative alla forma della distribuzione sottostante. Si presume che i dati rappresentino un campione casuale.

Per ottenere un test binomiale

- Dai menu, scegliere:
  Analizza > Test non parametrici > Finestre legacy > Binomiale...

Figura 27-51
Finestra di dialogo Test binomiale



- Selezionare una o più variabili numeriche oggetto del test.
- È possibile fare clic su Opzioni per ottenere statistiche descrittive, quartili e controllo delle modalità di elaborazione dei dati mancanti.

Test binomiale: Opzioni

Figura 27-52
Finestra di dialogo Test binomiale: Opzioni

Statistiche. È possibile scegliere uno o entrambe le statistiche riassuntive.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile da verificare verranno esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TEST (test binomiale)

Il linguaggio della sintassi dei comandi consente inoltre di:

- Selezionare gruppi specifici (ed escluderne altri) quando una variabile ha più di due categorie (con il sottocomando `BINOMIAL`).
- Specificare diversi punti di divisione o probabilità per variabili diverse (con il sottocomando `BINOMIAL`).
- Eseguire test confrontando la stessa variabile con diversi punti di divisione o probabilità (con il sottocomando `EXPECTED`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Test delle successioni

Il test delle successioni verifica se l'ordine delle occorrenze di due valori di una variabile è casuale. Una successione è una sequenza di osservazioni simili. Un campione con troppe o troppo poche successioni indica che il campione non è casuale.

Esempi. Si supponga che a venti persone venga chiesto se comprerebbero un determinato prodotto. La casualità prevista per il campione viene messa fortemente in dubbio se tutte le venti persone sono dello stesso sesso. È possibile utilizzare il test delle successioni per determinare se il campione è stato definito in modo casuale.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili.

Dati. Le variabili devono essere numeriche. Per convertire le variabili stringa in variabili numeriche, utilizzare il comando Ricodifica automatica del menu Trasforma.

Assunzioni. I test non parametrici non richiedono ipotesi relative alla forma della distribuzione sottostante. Utilizzare campioni da distribuzioni di probabilità continue.

Per ottenere un test delle successioni

▶ Dai menu, scegliere:
Analizza > Test non parametrici > Finestre legacy > Successioni...

Figura 27-53
Aggiunta di un punto di divisione personalizzato

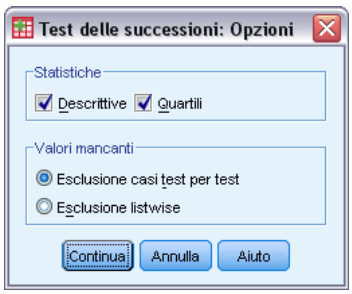
- Selezionare una o più variabili numeriche oggetto del test.
- È possibile fare clic su Opzioni per ottenere statistiche descrittive, quartili e controllo delle modalità di elaborazione dei dati mancanti.

Test delle successioni: Punto di divisione

Punto di divisione. Specifica un punto di divisione per dicotomizzare le variabili scelte dall'utente. È possibile utilizzare la media osservata, la mediana o la moda oppure un valore specificato come punto di divisione. I casi con valori minori del punto di divisione sono assegnati a un gruppo e i casi con valori uguali o maggiori del punto di divisione sono assegnati a un altro gruppo. Viene eseguito un test per ogni punto di divisione selezionato.

Opzioni test delle successioni

Figura 27-54
Finestra di dialogo Test delle successioni: Opzioni



Statistiche. È possibile scegliere uno o entrambe le statistiche riassuntive.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile sono esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TESTS (Test delle successioni)

Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare diversi punti di divisione per diverse variabili (con il sottocomando `RUNS`).
- Verificare la stessa variabile rispetto a diversi punti di divisione personalizzati (con il sottocomando `RUNS`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Test di Kolmogorov-Smirnov per un campione

La procedura Test di Kolmogorov-Smirnov per un campione consente di confrontare la funzione di distribuzione cumulata osservata per una variabile con la distribuzione teorica specificata, che può essere normale, uniforme o di Poisson. La *Z* di Kolmogorov-Smirnov viene calcolata in base alla differenza maggiore (in valore assoluto) tra la funzione di distribuzione cumulata osservata e teorica. Questo test sulla bontà di adattamento permette di verificare se le osservazioni possono provenire dalla distribuzione specificata.

Esempio. Molti test parametrici richiedono variabili distribuite in modo normale. Il test di Kolmogorov-Smirnov per un campione può essere utilizzato per verificare che una variabile, ad esempio *reddito*, sia distribuita in modo normale.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili.

Dati. Utilizzare variabili quantitative (misurazione a livello di intervallo o di rapporto).

Assunzioni. Il test di Kolmogorov-Smirnov presume che i parametri della distribuzione del test vengano specificati anticipatamente. Questa procedura consente di valutare i parametri del campione. La media e la deviazione standard del campione sono i parametri della distribuzione normale, i valori minimo e massimo del campione definiscono l'intervallo di distribuzione uniforme e la media del campione è il parametro per la distribuzione Poisson e per la distribuzione esponenziale. La capacità del test di rilevare gli scostamenti dalla distribuzione ipotizzata può essere seriamente compromessa. Per effettuare test su una distribuzione normale con parametri stimati, è generalmente consigliabile usare il test K-S di Lilliefors corretto (selezionabile dalla procedura Esplora).

Per ottenere un test di Kolmogorov-Smirnov per un campione

▶ Dai menu, scegliere:
Analizza > Test non parametrici > Finestre legacy > K-S per 1 campione...

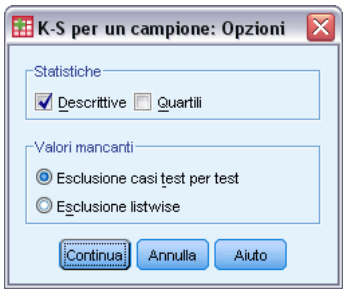
Figura 27-55
Finestra di dialogo Test di Kolmogorov-Smirnov per un campione

▶ Selezionare una o più variabili numeriche oggetto del test. Ogni variabile produce un test distinto.

▶ È possibile fare clic su Opzioni per ottenere statistiche descrittive, quartili e controllo delle modalità di elaborazione dei dati mancanti.

Test di Kolmogorov-Smirnov per un campione: Opzioni

Figura 27-56
Finestra di dialogo K-S per un campione



Statistiche. È possibile scegliere uno o entrambe le statistiche riassuntive.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile sono esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TESTS (Test di Kolmogorov-Smirnov per un campione)

Il linguaggio della sintassi dei comandi consente anche di specificare i parametri della distribuzione del test (con il sottocomando `K-S`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Test per due campioni indipendenti

La procedura del test per due campioni indipendenti consente di confrontare due gruppi di casi in base a una sola variabile.

Esempio. Sono stati creati nuovi apparecchi odontoiatrici che presentano numerosi vantaggi in termini di comodità, estetica ed efficacia ai fini dell'allineamento dei denti. Per determinare se i nuovi e i vecchi apparecchi devono essere portati per lo stesso periodo, sono stati scelti casualmente 10 bambini con il vecchio apparecchio e 10 bambini con il nuovo apparecchio. Dal test *U* di Mann-Whitney è possibile riscontrare che in media il nuovo apparecchio deve essere portato per un periodo di tempo inferiore.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili. Test: test *U* di Mann-Whitney, reazioni estreme di Moses, test *Z* di Kolmogorov-Smirnov, test delle successioni di Wald-Wolfowitz.

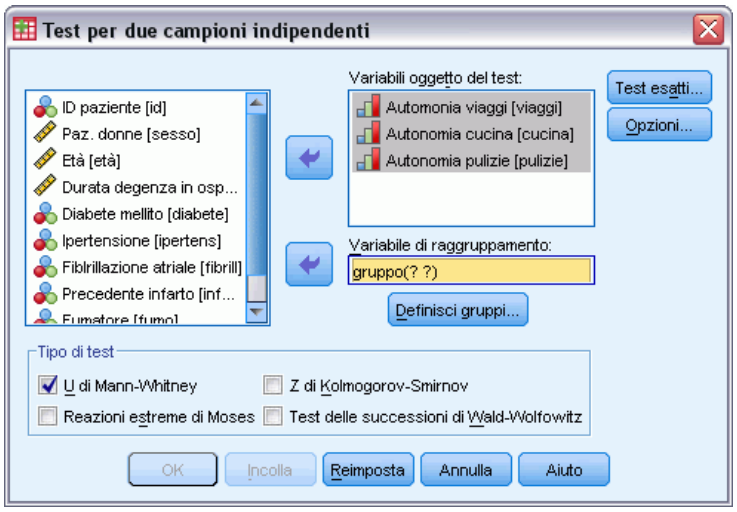
Dati. Utilizzare variabili numeriche che possono essere ordinate.

Assunzioni. Utilizzare campioni casuali indipendenti. Il test *U* di Mann-Whitney verifica l'uguaglianza di due distribuzioni. Per poterlo utilizzare per verificare le differenze nell'ubicazione di due distribuzioni, è necessario presupporre che le distribuzioni abbiano la stessa forma.

Per ottenere un test per due campioni indipendenti

▶ Dai menu, scegliere:
Analizza > Test non parametrici > Finestre legacy > 2 campioni indipendenti...

Figura 27-57
Finestra di dialogo Test per due campioni indipendenti



- Selezionare una o più variabili numeriche.
- Selezionare una variabile di raggruppamento e fare clic su Definisci gruppi per suddividere il file in due gruppi o campioni.

Tipi di test per due campioni indipendenti

Tipo di test. Sono disponibili quattro test che consentono di verificare se due campioni (gruppi) indipendenti provengono dalla stessa popolazione.

Il **test U di Mann-Whitney** è il test per due campioni indipendenti più diffuso. Equivale al test di Wilcoxon e al test di Kruskal-Wallis per due gruppi. Il test di Mann-Whitney permette di verificare l'equivalenza della posizione delle due popolazioni campione. Le osservazioni di entrambi i gruppi vengono combinate e ordinate per ranghi, assegnando la media dei ranghi ai valori a pari merito. Il numero di valori a pari merito deve essere inferiore al numero totale di osservazioni. Se la posizione delle popolazioni risulta identica, è necessario distribuire casualmente i ranghi tra i due campioni. Il test calcola il numero di volte in cui il punteggio del gruppo 1 è inferiore a quello del gruppo 2 e il numero di volte che un punteggio del gruppo 2 è inferiore al punteggio del gruppo 1. Il dato statistico che si ottiene con il test *U* di Mann-Whitney è inferiore a questi due numeri. Viene visualizzata anche la statistica *W* della somma dei ranghi di Wilcoxon. *W* è la somma dei ranghi del gruppo con il rango medio minore, a meno che i gruppi non abbiano lo stesso rango medio: in tal caso, è la somma dei ranghi del gruppo indicato per ultimo nella finestra di dialogo Due campioni indipendenti: Definisci gruppi.

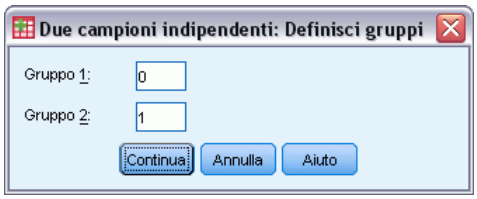
Il **test Z di Kolmogorov-Smirnov** e il **test delle successioni di Wald-Wolfowitz** sono test di carattere più generale che consentono di individuare le differenze tra le distribuzioni in termini di forma e posizione. Il test di Kolmogorov-Smirnov si basa sulla massima differenza in valore assoluto tra le funzioni di distribuzione cumulative osservate per entrambi i campioni. Quando tale differenza è significativa, le due distribuzioni vengono considerate diverse. Il test delle successioni di Wald-Wolfowitz consente di combinare e ordinare in ranghi le osservazioni di

entrambi i gruppi. Se i due campioni provengono dalla stessa popolazione, è necessario distribuire casualmente i gruppi all'interno della classifica.

Il **test delle reazioni estreme di Moses** si basa sull'ipotesi che la variabile sperimentale influenzi alcuni soggetti in una direzione e altri nella direzione opposta. Consente di verificare le risposte estreme confrontandole con un gruppo di controllo. Questo test è incentrato sull'estensione del gruppo di controllo e definisce la misura in cui i valori estremi del gruppo sperimentale influenzano l'estensione in caso di combinazione con il gruppo di controllo. Il gruppo di controllo viene definito dal valore del gruppo 1 nella finestra di dialogo Due campioni indipendenti: Definisci gruppi. Le osservazioni eseguite su entrambi i gruppi vengono combinate e classificate per ranghi. L'estensione del gruppo di controllo viene calcolata come la differenza tra i ranghi dei valori massimo e minimo del gruppo di controllo più 1. Poiché a causa di valori casuali anomali è probabile che l'intervallo dell'estensione risulti distorto, da ciascun estremo viene eliminato il 5% dei casi di controllo.

Test per due campioni indipendenti: Definisci gruppi

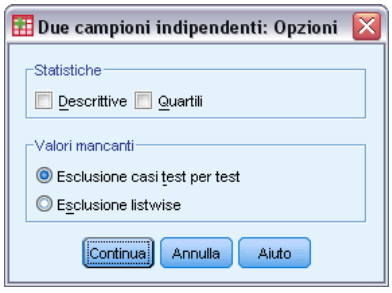
Figura 27-58
Finestra di dialogo Due campioni indipendenti: Definisci gruppi



Per suddividere il file in due gruppi o campioni, inserire un valore intero per il gruppo 1 e un altro per il gruppo 2. I casi con altri valori verranno esclusi dall'analisi.

Test per due campioni indipendenti: Opzioni

Figura 27-59
Finestra di dialogo Due campioni indipendenti: Opzioni



Statistiche. È possibile scegliere uno o entrambe le statistiche riassuntive.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile sono esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TESTS (Due campioni indipendenti)

Il linguaggio della sintassi dei comandi permette anche di specificare il numero di casi da eliminare dal test di Moses (con il sottocomando `MOSES`).

Vedere *Command Syntax Reference* per informazioni dettagliate sulla sintassi.

Test per due campioni dipendenti

La procedura del test per due campioni dipendenti consente di confrontare la distribuzione di due variabili.

Esempio. In generale, le famiglie ricevono l'intero prezzo di offerta per la vendita della propria casa? Applicando il test di Wilcoxon ai dati relativi a 10 case, si riscontrerà che sette famiglie ricevono una somma inferiore al prezzo di offerta, una famiglia riceve una somma superiore, mentre due sole famiglie lo ricevono interamente.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili. Test: di Wilcoxon, del segno e di McNemar. Se è installata l'opzione Test esatti (disponibile solo nei sistemi operativi Windows), è disponibile anche il test di omogeneità marginale.

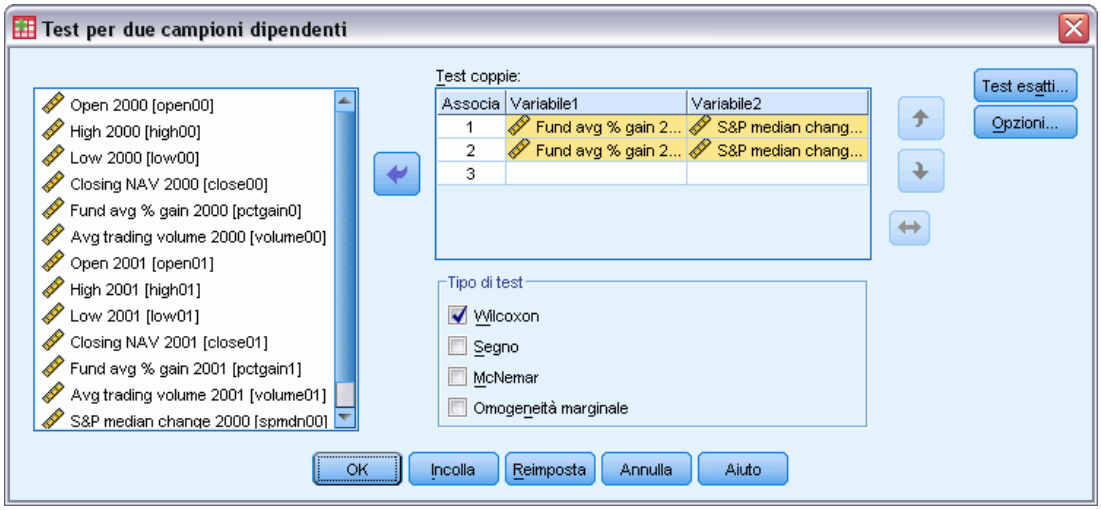
Dati. Utilizzare variabili numeriche che possono essere ordinate.

Assunzioni. Anche se per le due variabili non si ipotizza una particolare distribuzione, si presume che la distribuzione delle differenze a coppie sia simmetrica.

Per ottenere test per due campioni dipendenti

▶ Dai menu, scegliere:
Analizza > Test non parametrici > Finestre legacy > 2 campioni dipendenti...

Figura 27-60
Finestra di dialogo Test per due campioni dipendenti



▶ Selezionare una o più coppie di variabili.

Tipi di test per due campioni dipendenti

I test descritti in questa sezione permettono di confrontare le distribuzioni di due variabili correlate. Il test più appropriato varia a seconda dei tipi di dati.

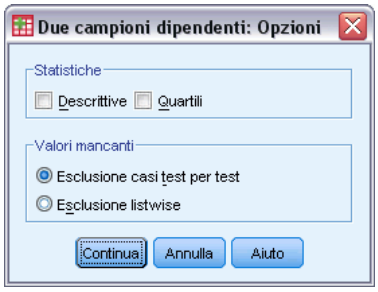
Se i dati sono continui, utilizzare il test del segno o di Wilcoxon. Il **test del segno** permette di calcolare le differenze tra le due variabili per tutti i casi e di classificarle come positive, negative o a pari merito. Se le due variabili sono distribuite in modo analogo, il numero di differenze positive e negative non differirà in misura significativa. Il **test di Wilcoxon** prende in considerazione le informazioni relative al segno e all'entità delle differenze tra le coppie. Poiché il test di Wilcoxon include un maggior numero di informazioni relative ai dati, risulta più valido del test del segno.

Se i dati sono binari, utilizzare il **test di McNemar**. Questo test viene in genere utilizzato in presenza di misure ripetute, ovvero quando la risposta del soggetto viene richiesta due volte: prima e dopo il verificarsi di un determinato evento. Il test di McNemar consente di determinare se il tasso di risposta iniziale (prima dell'evento) equivale al tasso di risposta finale (dopo l'evento). Questo test risulta particolarmente utile per individuare le variazioni della risposta in disegni sperimentali del tipo 'prima e dopo'.

Se i dati sono categoriali, utilizzare il **test di omogeneità marginale**. Estensione del test di McNemar dalla risposta binaria a quella multinomiale. Consente di verificare le variazioni della risposta utilizzando la distribuzione del chi-quadrato e risulta utile in disegni sperimentali del tipo 'prima e dopo'. Il test di omogeneità marginale è disponibile solo se è stato installato il modulo Exact Tests.

Test per due campioni dipendenti: Opzioni

Figura 27-61
Finestra di dialogo Due campioni dipendenti: Opzioni



Statistiche. È possibile scegliere uno o entrambe le statistiche riassuntive.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile sono esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TESTS (due campioni dipendenti)

Il linguaggio della sintassi dei comandi permette anche di verificare una variabile con ciascuna variabile dell'elenco.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Test per diversi campioni indipendenti

La procedura per i test per diversi campioni indipendenti consente di confrontare due o più gruppi di casi in base a una variabile.

Esempio. Le lampadine da 100 watt di tre diversi produttori si differenziano in relazione al tempo medio di bruciatura del filamento? Grazie all'ANOVA univariata di Kruskal-Wallis è possibile verificare che la durata media delle tre lampadine è effettivamente diversa.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili. Test: *H* di Kruskal-Wallis, mediana.

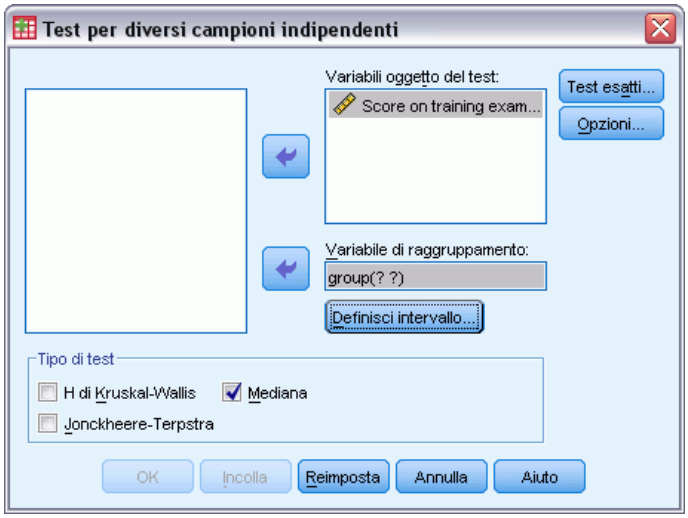
Dati. Utilizzare variabili numeriche che possono essere ordinate.

Assunzioni. Utilizzare campioni casuali indipendenti. IL test *H* di Kruskal-Wallis richiede che i campioni sottoposti a test siano simili per forma.

Per ottenere test per diversi campioni indipendenti

- Dai menu, scegliere:
  Analizza > Test non parametrici > Finestre legacy > K campioni indipendenti...

Figura 27-62
Definizione del test della mediana



- Selezionare una o più variabili numeriche.

- Selezionare una variabile di raggruppamento e fare clic su Definisci intervallo per specificare i valori interi minimo e massimo per la variabile di raggruppamento.

Test per diversi campioni indipendenti: tipi di test

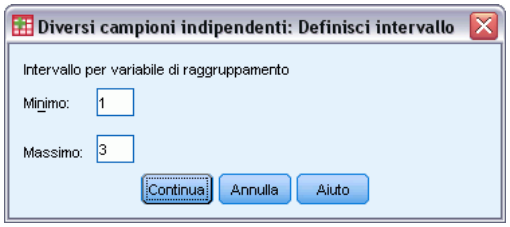
Sono disponibili tre test per stabilire se diversi campioni indipendenti sono stati estratti dalla stessa popolazione. Il test *H* di Kruskal-Wallis, il test della mediana e il test di Jonckheere-Terpstra consentono di verificare se i diversi campioni indipendenti sono stati estratti dalla stessa popolazione.

Il test **H di Kruskal-Wallis**, un'estensione del test *U* di Mann-Whitney, è la versione non parametrica dell'analisi univariata della varianza e consente di rilevare le differenze nella posizione di distribuzione. Il **test della mediana**, che è più generale ma non altrettanto potente, consente di rilevare le differenze distribuzionali nella posizione e nella forma. Il test *H* di Kruskal-Wallis e il test della mediana presumono che non esistano ordinamenti *a priori* delle *k* popolazioni da cui sono estratti i campioni.

Quando *esiste* un naturale ordinamento *a priori* (crescente o decrescente) delle *k* popolazioni, il **test di Jonckheere-Terpstra** è più potente. Ad esempio, le *k* popolazioni possono rappresentare *k* temperature crescenti. L'ipotesi che diverse temperature producano la stessa distribuzione della risposta è verificata rispetto all'ipotesi alternativa in base a cui al salire della temperatura, cresce il valore della risposta. Qui l'ipotesi alternativa è ordinata e quindi il test di Jonckheere-Terpstra è il più appropriato da utilizzare. Il test di Jonckheere-Terpstra è disponibile solo se è installato il modulo aggiuntivo Testi esatti.

Test per diversi campioni indipendenti: Definisci intervallo

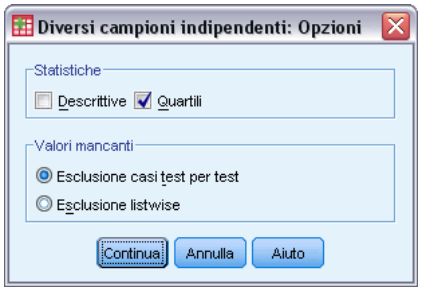
Figura 27-63
Finestra di dialogo Diversi campioni indipendenti: Definisci intervallo



Per definire l'intervallo, immettere valori interni per il minimo e il massimo che corrispondono alle categorie minore e maggiore della variabile di raggruppamento. Sono esclusi i casi con valori al di fuori dei limiti. Se, ad esempio, si specifica un limite inferiore di 1 e un limite superiore di 3, verranno utilizzati solo i valori interi compresi tra 1 e 3. Il valore minimo deve essere inferiore al valore massimo ed entrambi i valori devono essere specificati.

Test per diversi campioni indipendenti: Opzioni

Figura 27-64
Finestra di dialogo Diversi campioni indipendenti: Opzioni



Statistiche. È possibile scegliere uno o entrambe le statistiche riassuntive.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile sono esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TESTS (K campioni indipendenti)

Il linguaggio della sintassi dei comandi permette anche di specificare un valore diverso dalla mediana osservata per il test della mediana (con il sottocomando `MEDIAN`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Test per diversi campioni dipendenti

Il test per diversi campioni dipendenti consente di confrontare le distribuzioni di due o più variabili.

Esempio. Il pubblico associa diversi livello di prestigio al ruolo di dottore, avvocato, ufficiale della polizia e insegnante? A dieci persone viene chiesto di ordinare queste quattro occupazioni in base al prestigio. Il test di Friedman indica che il pubblico associa effettivamente livelli di prestigio diversi a queste quattro professioni.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili. Test: Friedman, *W* di Kendall e *Q* di Cochran.

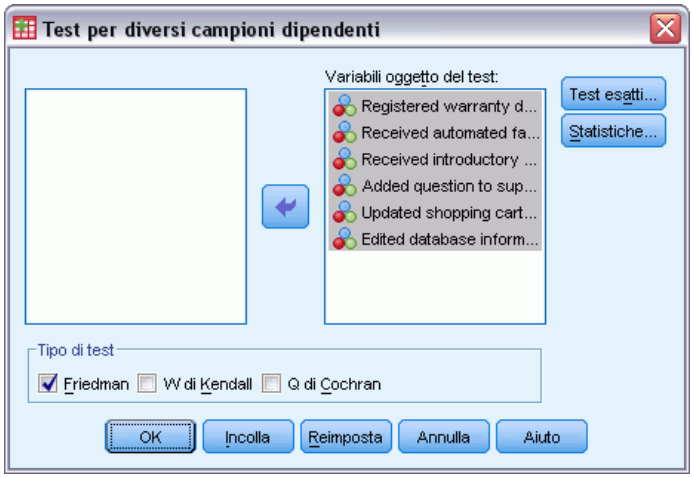
Dati. Utilizzare variabili numeriche che possono essere ordinate.

Assunzioni. I test non parametrici non richiedono ipotesi relative alla forma della distribuzione sottostante. Utilizzare campioni casuali dipendenti.

Per ottenere i test per diversi campioni dipendenti

▶ Dai menu, scegliere:
Analizza > Test non parametrici > Finestre legacy > K campioni dipendenti...

Figura 27-65
Selezione di Cochran come tipo di test



▶ Selezionare una o più variabili oggetto del test numeriche.

Test per diversi campioni dipendenti: tipi di test

Sono disponibili tre test per confrontare le distribuzioni di diverse variabili correlate.

Il **testdi Friedman** è l'equivalente non parametrico di un disegno di misure ripetute per un campione o ANOVA a due vie con una osservazione per cella. Friedman verifica l'ipotesi nulla secondo cui *k* variabili correlate provengono dalla stessa popolazione. Per ogni caso, alle *k* variabili viene assegnato un rango da 1 a *k*. Le statistiche del test sono basate su questi ranghi.

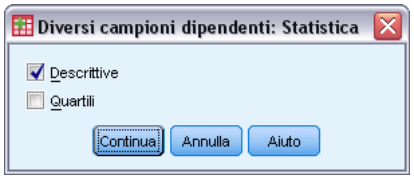
W di Kendall è una normalizzazione delle statistiche di Friedman. È possibile interpretare il *W* di Kendall come il coefficiente di concordanza, che rappresenta la misura dell'accordo tra stimatori. Ogni caso è uno stimatore e ogni variabile è un elemento o individuo da stimare. Per

ogni variabile viene calcolata la somma dei ranghi. *W* di Kendall varia tra 0 (nessun accordo) e 1 (accordo completo).

Q di Cochran è identico al test di Friedman ma è applicabile quando tutte le risposte sono binarie. Questo test è un'estensione del test di McNemar alla situazione di *k*-campioni. I test *Q* di Cochran verificano l'ipotesi secondo cui diverse variabili dicotomiche hanno la stessa media. Le variabili sono misurate sullo stesso individuo o su individui collegati fra loro.

Test per diversi campioni dipendenti: Statistica

Figura 27-66
Finestra di dialogo Diversi campioni dipendenti: Statistica



È possibile scegliere le statistiche.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Funzioni aggiuntive del comando NPAR TESTS (K campioni dipendenti)

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Test binomiale

Grazie alla procedura del test binomiale è possibile confrontare le frequenze osservate delle due categorie di una variabile dicotomica con le frequenze previste in presenza di una distribuzione binomiale con il parametro di probabilità specificato. Per impostazione predefinita, il parametro di probabilità per entrambi i gruppi è 0,5. Per modificare le probabilità, è possibile inserire una proporzione di prova per il primo gruppo. La probabilità per il secondo gruppo sarà uguale a 1 meno la probabilità specificata per il primo gruppo.

Esempio. Quando si lancia in aria una moneta, la probabilità che esca testa è pari a 1/2. In base a questa ipotesi, la moneta viene lanciata in aria 40 volte e i risultati (testa o croce) vengono registrati. Dal test binomiale può risultare che per i 3/4 dei lanci della moneta è uscita testa e che il livello di significatività è molto basso (0,0027). Questi risultati indicano che la probabilità che venga testa molto spesso non è pari a 1/2; pertanto, la stima sul comportamento della moneta probabilmente risulta distorta.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili.

Dati. Le variabili incluse nel test devono essere numeriche e dicotomiche. Per convertire le variabili stringa in variabili numeriche, utilizzare il comando Ricodifica automatica del menu Trasforma. Una **variabile dicotomica** è una variabile che prevede solo due possibili valori: *Sì* o

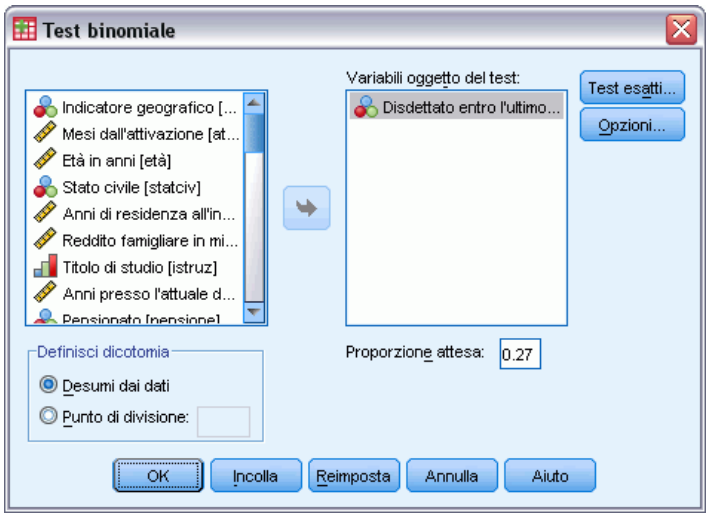
No, *Vero* o *Falso*, 0 o 1 e così via. Il primo valore rilevato nell'insieme di dati definisce il primo gruppo, mentre l'altro valore definisce il secondo gruppo. Se le variabili sono dicotomiche, è necessario specificare un punto di divisione. Utilizzando il punto di divisione è possibile assegnare al primo gruppo i casi con valori inferiori o uguali al punto di divisione e i rimanenti casi al secondo gruppo.

Assunzioni. I test non parametrici non richiedono ipotesi relative alla forma della distribuzione sottostante. Si presume che i dati rappresentino un campione casuale.

Per ottenere un test binomiale

- Dai menu, scegliere:
  Analizza > Test non parametrici > Finestre legacy > Binomiale...

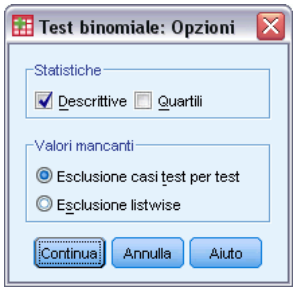
Figura 27-67
Finestra di dialogo Test binomiale



- Selezionare una o più variabili numeriche oggetto del test.
- È possibile fare clic su Opzioni per ottenere statistiche descrittive, quartili e controllo delle modalità di elaborazione dei dati mancanti.

Test binomiale: Opzioni

Figura 27-68
Finestra di dialogo Test binomiale: Opzioni



Statistiche. È possibile scegliere uno o entrambe le statistiche riassuntive.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile da verificare verranno esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TEST (test binomiale)

Il linguaggio della sintassi dei comandi consente inoltre di:

- Selezionare gruppi specifici (ed escluderne altri) quando una variabile ha più di due categorie (con il sottocomando `BINOMIAL`).
- Specificare diversi punti di divisione o probabilità per variabili diverse (con il sottocomando `BINOMIAL`).
- Eseguire test confrontando la stessa variabile con diversi punti di divisione o probabilità (con il sottocomando `EXPECTED`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Test delle successioni

Il test delle successioni verifica se l'ordine delle occorrenze di due valori di una variabile è casuale. Una successione è una sequenza di osservazioni simili. Un campione con troppe o troppo poche successioni indica che il campione non è casuale.

Esempi. Si supponga che a venti persone venga chiesto se comprerebbero un determinato prodotto. La casualità prevista per il campione viene messa fortemente in dubbio se tutte le venti persone sono dello stesso sesso. È possibile utilizzare il test delle successioni per determinare se il campione è stato definito in modo casuale.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili.

Dati. Le variabili devono essere numeriche. Per convertire le variabili stringa in variabili numeriche, utilizzare il comando Ricodifica automatica del menu Trasforma.

Assunzioni. I test non parametrici non richiedono ipotesi relative alla forma della distribuzione sottostante. Utilizzare campioni da distribuzioni di probabilità continue.

Per ottenere un test delle successioni

▶ Dai menu, scegliere:
Analizza > Test non parametrici > Finestre legacy > Successioni...

Figura 27-69
Aggiunta di un punto di divisione personalizzato

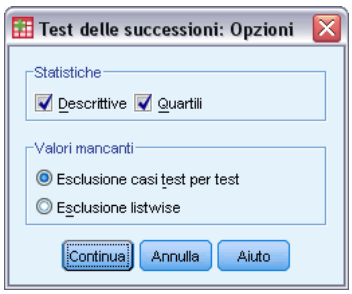
- Selezionare una o più variabili numeriche oggetto del test.
- È possibile fare clic su Opzioni per ottenere statistiche descrittive, quartili e controllo delle modalità di elaborazione dei dati mancanti.

Test delle successioni: Punto di divisione

Punto di divisione. Specifica un punto di divisione per dicotomizzare le variabili scelte dall'utente. È possibile utilizzare la media osservata, la mediana o la moda oppure un valore specificato come punto di divisione. I casi con valori minori del punto di divisione sono assegnati a un gruppo e i casi con valori uguali o maggiori del punto di divisione sono assegnati a un altro gruppo. Viene eseguito un test per ogni punto di divisione selezionato.

Opzioni test delle successioni

Figura 27-70
Finestra di dialogo Test delle successioni: Opzioni



Statistiche. È possibile scegliere uno o entrambe le statistiche riassuntive.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile sono esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TESTS (Test delle successioni)

Il linguaggio della sintassi dei comandi consente inoltre di:

- Specificare diversi punti di divisione per diverse variabili (con il sottocomando `RUNS`).
- Verificare la stessa variabile rispetto a diversi punti di divisione personalizzati (con il sottocomando `RUNS`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Test di Kolmogorov-Smirnov per un campione

La procedura Test di Kolmogorov-Smirnov per un campione consente di confrontare la funzione di distribuzione cumulata osservata per una variabile con la distribuzione teorica specificata, che può essere normale, uniforme o di Poisson. La *Z* di Kolmogorov-Smirnov viene calcolata in base alla differenza maggiore (in valore assoluto) tra la funzione di distribuzione cumulata osservata e teorica. Questo test sulla bontà di adattamento permette di verificare se le osservazioni possono provenire dalla distribuzione specificata.

Esempio. Molti test parametrici richiedono variabili distribuite in modo normale. Il test di Kolmogorov-Smirnov per un campione può essere utilizzato per verificare che una variabile, ad esempio *reddito*, sia distribuita in modo normale.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili.

Dati. Utilizzare variabili quantitative (misurazione a livello di intervallo o di rapporto).

Assunzioni. Il test di Kolmogorov-Smirnov presume che i parametri della distribuzione del test vengano specificati anticipatamente. Questa procedura consente di valutare i parametri del campione. La media e la deviazione standard del campione sono i parametri della distribuzione normale, i valori minimo e massimo del campione definiscono l'intervallo di distribuzione uniforme e la media del campione è il parametro per la distribuzione Poisson e per la distribuzione esponenziale. La capacità del test di rilevare gli scostamenti dalla distribuzione ipotizzata può essere seriamente compromessa. Per effettuare test su una distribuzione normale con parametri stimati, è generalmente consigliabile usare il test K-S di Lilliefors corretto (selezionabile dalla procedura Esplora).

Per ottenere un test di Kolmogorov-Smirnov per un campione

▶ Dai menu, scegliere:
Analizza > Test non parametrici > Finestre legacy > K-S per 1 campione...

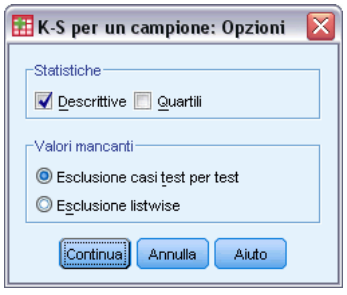
Figura 27-71
Finestra di dialogo Test di Kolmogorov-Smirnov per un campione

▶ Selezionare una o più variabili numeriche oggetto del test. Ogni variabile produce un test distinto.

▶ È possibile fare clic su Opzioni per ottenere statistiche descrittive, quartili e controllo delle modalità di elaborazione dei dati mancanti.

Test di Kolmogorov-Smirnov per un campione: Opzioni

Figura 27-72
Finestra di dialogo K-S per un campione



Statistiche. È possibile scegliere uno o entrambe le statistiche riassuntive.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile sono esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TESTS (Test di Kolmogorov-Smirnov per un campione)

Il linguaggio della sintassi dei comandi consente anche di specificare i parametri della distribuzione del test (con il sottocomando `K-S`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Test per due campioni indipendenti

La procedura del test per due campioni indipendenti consente di confrontare due gruppi di casi in base a una sola variabile.

Esempio. Sono stati creati nuovi apparecchi odontoiatrici che presentano numerosi vantaggi in termini di comodità, estetica ed efficacia ai fini dell'allineamento dei denti. Per determinare se i nuovi e i vecchi apparecchi devono essere portati per lo stesso periodo, sono stati scelti casualmente 10 bambini con il vecchio apparecchio e 10 bambini con il nuovo apparecchio. Dal test *U* di Mann-Whitney è possibile riscontrare che in media il nuovo apparecchio deve essere portato per un periodo di tempo inferiore.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili. Test: test *U* di Mann-Whitney, reazioni estreme di Moses, test *Z* di Kolmogorov-Smirnov, test delle successioni di Wald-Wolfowitz.

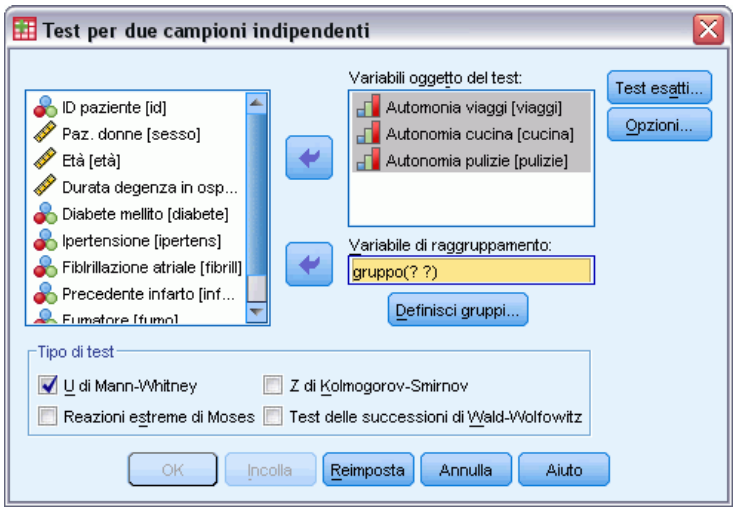
Dati. Utilizzare variabili numeriche che possono essere ordinate.

Assunzioni. Utilizzare campioni casuali indipendenti. Il test *U* di Mann-Whitney verifica l'uguaglianza di due distribuzioni. Per poterlo utilizzare per verificare le differenze nell'ubicazione di due distribuzioni, è necessario presupporre che le distribuzioni abbiano la stessa forma.

Per ottenere un test per due campioni indipendenti

▶ Dai menu, scegliere:
Analizza > Test non parametrici > Finestre legacy > 2 campioni indipendenti...

Figura 27-73
Finestra di dialogo Test per due campioni indipendenti



- Selezionare una o più variabili numeriche.
- Selezionare una variabile di raggruppamento e fare clic su Definisci gruppi per suddividere il file in due gruppi o campioni.

Tipi di test per due campioni indipendenti

Tipo di test. Sono disponibili quattro test che consentono di verificare se due campioni (gruppi) indipendenti provengono dalla stessa popolazione.

Il **test U di Mann-Whitney** è il test per due campioni indipendenti più diffuso. Equivale al test di Wilcoxon e al test di Kruskal-Wallis per due gruppi. Il test di Mann-Whitney permette di verificare l'equivalenza della posizione delle due popolazioni campione. Le osservazioni di entrambi i gruppi vengono combinate e ordinate per ranghi, assegnando la media dei ranghi ai valori a pari merito. Il numero di valori a pari merito deve essere inferiore al numero totale di osservazioni. Se la posizione delle popolazioni risulta identica, è necessario distribuire casualmente i ranghi tra i due campioni. Il test calcola il numero di volte in cui il punteggio del gruppo 1 è inferiore a quello del gruppo 2 e il numero di volte che un punteggio del gruppo 2 è inferiore al punteggio del gruppo 1. Il dato statistico che si ottiene con il test *U* di Mann-Whitney è inferiore a questi due numeri. Viene visualizzata anche la statistica *W* della somma dei ranghi di Wilcoxon. *W* è la somma dei ranghi del gruppo con il rango medio minore, a meno che i gruppi non abbiano lo stesso rango medio: in tal caso, è la somma dei ranghi del gruppo indicato per ultimo nella finestra di dialogo Due campioni indipendenti: Definisci gruppi.

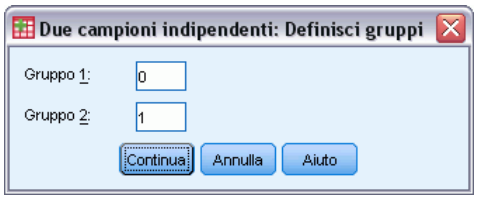
Il **test Z di Kolmogorov-Smirnov** e il **test delle successioni di Wald-Wolfowitz** sono test di carattere più generale che consentono di individuare le differenze tra le distribuzioni in termini di forma e posizione. Il test di Kolmogorov-Smirnov si basa sulla massima differenza in valore assoluto tra le funzioni di distribuzione cumulative osservate per entrambi i campioni. Quando tale differenza è significativa, le due distribuzioni vengono considerate diverse. Il test delle successioni di Wald-Wolfowitz consente di combinare e ordinare in ranghi le osservazioni di

entrambi i gruppi. Se i due campioni provengono dalla stessa popolazione, è necessario distribuire casualmente i gruppi all'interno della classifica.

Il **test delle reazioni estreme di Moses** si basa sull'ipotesi che la variabile sperimentale influenzi alcuni soggetti in una direzione e altri nella direzione opposta. Consente di verificare le risposte estreme confrontandole con un gruppo di controllo. Questo test è incentrato sull'estensione del gruppo di controllo e definisce la misura in cui i valori estremi del gruppo sperimentale influenzano l'estensione in caso di combinazione con il gruppo di controllo. Il gruppo di controllo viene definito dal valore del gruppo 1 nella finestra di dialogo Due campioni indipendenti: Definisci gruppi. Le osservazioni eseguite su entrambi i gruppi vengono combinate e classificate per ranghi. L'estensione del gruppo di controllo viene calcolata come la differenza tra i ranghi dei valori massimo e minimo del gruppo di controllo più 1. Poiché a causa di valori casuali anomali è probabile che l'intervallo dell'estensione risulti distorto, da ciascun estremo viene eliminato il 5% dei casi di controllo.

Test per due campioni indipendenti: Definisci gruppi

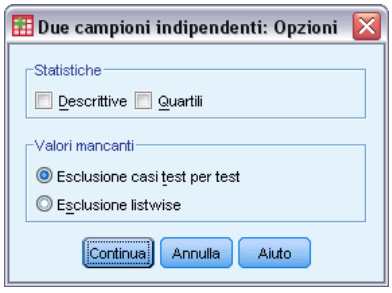
Figura 27-74
Finestra di dialogo Due campioni indipendenti: Definisci gruppi



Per suddividere il file in due gruppi o campioni, inserire un valore intero per il gruppo 1 e un altro per il gruppo 2. I casi con altri valori verranno esclusi dall'analisi.

Test per due campioni indipendenti: Opzioni

Figura 27-75
Finestra di dialogo Due campioni indipendenti: Opzioni



Statistiche. È possibile scegliere uno o entrambe le statistiche riassuntive.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile sono esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TESTS (Due campioni indipendenti)

Il linguaggio della sintassi dei comandi permette anche di specificare il numero di casi da eliminare dal test di Moses (con il sottocomando `MOSES`).

Vedere *Command Syntax Reference* per informazioni dettagliate sulla sintassi.

Test per due campioni dipendenti

La procedura del test per due campioni dipendenti consente di confrontare la distribuzione di due variabili.

Esempio. In generale, le famiglie ricevono l'intero prezzo di offerta per la vendita della propria casa? Applicando il test di Wilcoxon ai dati relativi a 10 case, si riscontrerà che sette famiglie ricevono una somma inferiore al prezzo di offerta, una famiglia riceve una somma superiore, mentre due sole famiglie lo ricevono interamente.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili. Test: di Wilcoxon, del segno e di McNemar. Se è installata l'opzione Test esatti (disponibile solo nei sistemi operativi Windows), è disponibile anche il test di omogeneità marginale.

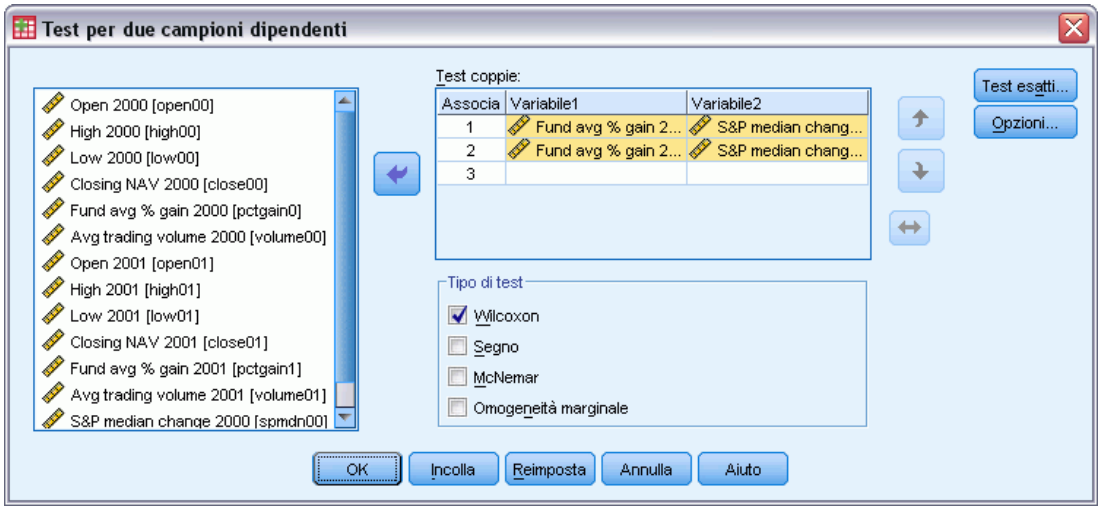
Dati. Utilizzare variabili numeriche che possono essere ordinate.

Assunzioni. Anche se per le due variabili non si ipotizza una particolare distribuzione, si presume che la distribuzione delle differenze a coppie sia simmetrica.

Per ottenere test per due campioni dipendenti

▶ Dai menu, scegliere:
Analizza > Test non parametrici > Finestre legacy > 2 campioni dipendenti...

Figura 27-76
Finestra di dialogo Test per due campioni dipendenti



▶ Selezionare una o più coppie di variabili.

Tipi di test per due campioni dipendenti

I test descritti in questa sezione permettono di confrontare le distribuzioni di due variabili correlate. Il test più appropriato varia a seconda dei tipi di dati.

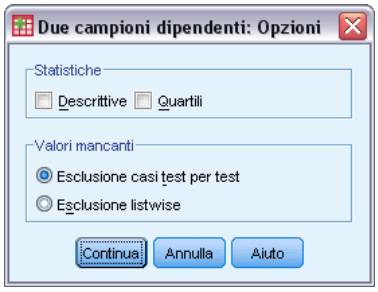
Se i dati sono continui, utilizzare il test del segno o di Wilcoxon. Il **test del segno** permette di calcolare le differenze tra le due variabili per tutti i casi e di classificarle come positive, negative o a pari merito. Se le due variabili sono distribuite in modo analogo, il numero di differenze positive e negative non differirà in misura significativa. Il **test di Wilcoxon** prende in considerazione le informazioni relative al segno e all'entità delle differenze tra le coppie. Poiché il test di Wilcoxon include un maggior numero di informazioni relative ai dati, risulta più valido del test del segno.

Se i dati sono binari, utilizzare il **test di McNemar**. Questo test viene in genere utilizzato in presenza di misure ripetute, ovvero quando la risposta del soggetto viene richiesta due volte: prima e dopo il verificarsi di un determinato evento. Il test di McNemar consente di determinare se il tasso di risposta iniziale (prima dell'evento) equivale al tasso di risposta finale (dopo l'evento). Questo test risulta particolarmente utile per individuare le variazioni della risposta in disegni sperimentali del tipo 'prima e dopo'.

Se i dati sono categoriali, utilizzare il **test di omogeneità marginale**. Estensione del test di McNemar dalla risposta binaria a quella multinomiale. Consente di verificare le variazioni della risposta utilizzando la distribuzione del chi-quadrato e risulta utile in disegni sperimentali del tipo 'prima e dopo'. Il test di omogeneità marginale è disponibile solo se è stato installato il modulo Exact Tests.

Test per due campioni dipendenti: Opzioni

Figura 27-77
Finestra di dialogo Due campioni dipendenti: Opzioni



Statistiche. È possibile scegliere uno o entrambe le statistiche riassuntive.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile sono esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TESTS (due campioni dipendenti)

Il linguaggio della sintassi dei comandi permette anche di verificare una variabile con ciascuna variabile dell'elenco.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Test per diversi campioni indipendenti

La procedura per i test per diversi campioni indipendenti consente di confrontare due o più gruppi di casi in base a una variabile.

Esempio. Le lampadine da 100 watt di tre diversi produttori si differenziano in relazione al tempo medio di bruciatura del filamento? Grazie all'ANOVA univariata di Kruskal-Wallis è possibile verificare che la durata media delle tre lampadine è effettivamente diversa.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili. Test: *H* di Kruskal-Wallis, mediana.

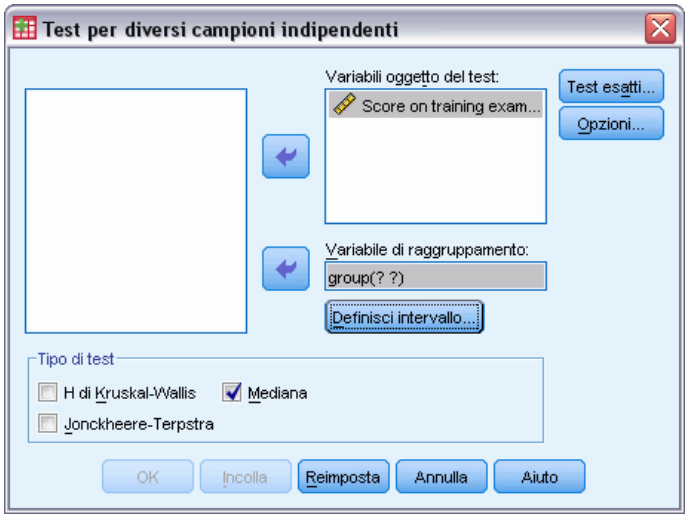
Dati. Utilizzare variabili numeriche che possono essere ordinate.

Assunzioni. Utilizzare campioni casuali indipendenti. IL test *H* di Kruskal-Wallis richiede che i campioni sottoposti a test siano simili per forma.

Per ottenere test per diversi campioni indipendenti

▶ Dai menu, scegliere:
Analizza > Test non parametrici > Finestre legacy > K campioni indipendenti...

Figura 27-78
Definizione del test della mediana



▶ Selezionare una o più variabili numeriche.

▶ Selezionare una variabile di raggruppamento e fare clic su Definisci intervallo per specificare i valori interi minimo e massimo per la variabile di raggruppamento.

Test per diversi campioni indipendenti: tipi di test

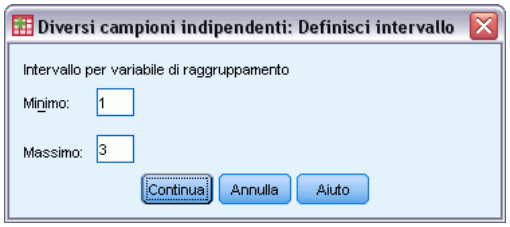
Sono disponibili tre test per stabilire se diversi campioni indipendenti sono stati estratti dalla stessa popolazione. Il test *H* di Kruskal-Wallis, il test della mediana e il test di Jonckheere-Terpstra consentono di verificare se i diversi campioni indipendenti sono stati estratti dalla stessa popolazione.

Il test **H di Kruskal-Wallis**, un'estensione del test *U* di Mann-Whitney, è la versione non parametrica dell'analisi univariata della varianza e consente di rilevare le differenze nella posizione di distribuzione. Il **test della mediana**, che è più generale ma non altrettanto potente, consente di rilevare le differenze distribuzionali nella posizione e nella forma. Il test *H* di Kruskal-Wallis e il test della mediana presumono che non esistano ordinamenti *a priori* delle *k* popolazioni da cui sono estratti i campioni.

Quando *esiste* un naturale ordinamento *a priori* (crescente o decrescente) delle *k* popolazioni, il **test di Jonckheere-Terpstra** è più potente. Ad esempio, le *k* popolazioni possono rappresentare *k* temperature crescenti. L'ipotesi che diverse temperature producano la stessa distribuzione della risposta è verificata rispetto all'ipotesi alternativa in base a cui al salire della temperatura, cresce il valore della risposta. Qui l'ipotesi alternativa è ordinata e quindi il test di Jonckheere-Terpstra è il più appropriato da utilizzare. Il test di Jonckheere-Terpstra è disponibile solo se è installato il modulo aggiuntivo Testi esatti.

Test per diversi campioni indipendenti: Definisci intervallo

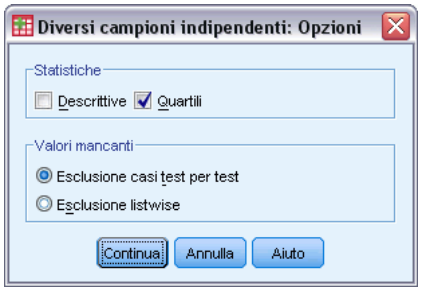
Figura 27-79
Finestra di dialogo Diversi campioni indipendenti: Definisci intervallo



Per definire l'intervallo, immettere valori interni per il minimo e il massimo che corrispondono alle categorie minore e maggiore della variabile di raggruppamento. Sono esclusi i casi con valori al di fuori dei limiti. Se, ad esempio, si specifica un limite inferiore di 1 e un limite superiore di 3, verranno utilizzati solo i valori interi compresi tra 1 e 3. Il valore minimo deve essere inferiore al valore massimo ed entrambi i valori devono essere specificati.

Test per diversi campioni indipendenti: Opzioni

Figura 27-80
Finestra di dialogo Diversi campioni indipendenti: Opzioni



Statistiche. È possibile scegliere uno o entrambe le statistiche riassuntive.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Valori mancanti. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Esclusione casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Esclusione listwise.** I casi con valori mancanti per qualsiasi variabile sono esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TESTS (K campioni indipendenti)

Il linguaggio della sintassi dei comandi permette anche di specificare un valore diverso dalla mediana osservata per il test della mediana (con il sottocomando `MEDIAN`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Test per diversi campioni dipendenti

Il test per diversi campioni dipendenti consente di confrontare le distribuzioni di due o più variabili.

Esempio. Il pubblico associa diversi livello di prestigio al ruolo di dottore, avvocato, ufficiale della polizia e insegnante? A dieci persone viene chiesto di ordinare queste quattro occupazioni in base al prestigio. Il test di Friedman indica che il pubblico associa effettivamente livelli di prestigio diversi a queste quattro professioni.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili. Test: Friedman, *W* di Kendall e *Q* di Cochran.

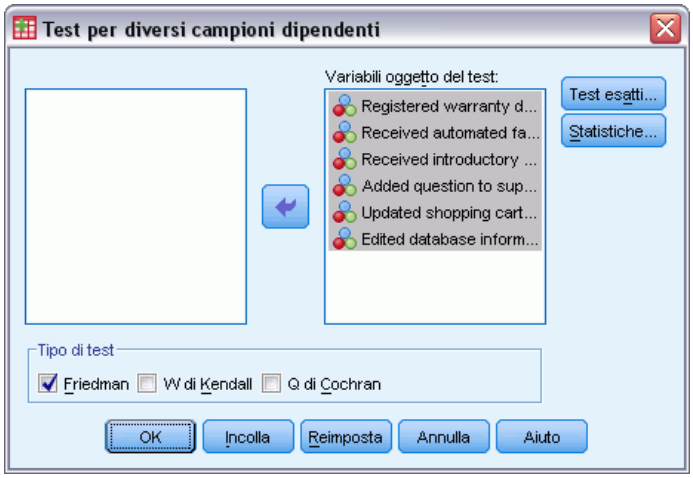
Dati. Utilizzare variabili numeriche che possono essere ordinate.

Assunzioni. I test non parametrici non richiedono ipotesi relative alla forma della distribuzione sottostante. Utilizzare campioni casuali dipendenti.

Per ottenere i test per diversi campioni dipendenti

▶ Dai menu, scegliere:
Analizza > Test non parametrici > Finestre legacy > K campioni dipendenti...

Figura 27-81
Selezione di Cochran come tipo di test



▶ Selezionare una o più variabili oggetto del test numeriche.

Test per diversi campioni dipendenti: tipi di test

Sono disponibili tre test per confrontare le distribuzioni di diverse variabili correlate.

Il **testdi Friedman** è l'equivalente non parametrico di un disegno di misure ripetute per un campione o ANOVA a due vie con una osservazione per cella. Friedman verifica l'ipotesi nulla secondo cui *k* variabili correlate provengono dalla stessa popolazione. Per ogni caso, alle *k* variabili viene assegnato un rango da 1 a *k*. Le statistiche del test sono basate su questi ranghi.

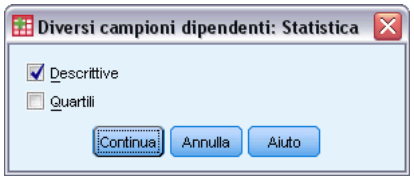
W di Kendall è una normalizzazione delle statistiche di Friedman. È possibile interpretare il *W* di Kendall come il coefficiente di concordanza, che rappresenta la misura dell'accordo tra stimatori. Ogni caso è uno stimatore e ogni variabile è un elemento o individuo da stimare. Per

ogni variabile viene calcolata la somma dei ranghi. *W* di Kendall varia tra 0 (nessun accordo) e 1 (accordo completo).

Q di Cochran è identico al test di Friedman ma è applicabile quando tutte le risposte sono binarie. Questo test è un'estensione del test di McNemar alla situazione di *k*-campioni. I test *Q* di Cochran verificano l'ipotesi secondo cui diverse variabili dicotomiche hanno la stessa media. Le variabili sono misurate sullo stesso individuo o su individui collegati fra loro.

Test per diversi campioni dipendenti: Statistica

Figura 27-82
Finestra di dialogo Diversi campioni dipendenti: Statistica



È possibile scegliere le statistiche.

- **Descrittive.** Consente di visualizzare la media, la deviazione standard, il valore minimo e massimo e il numero di casi non mancanti.
- **Quartili.** Consente di visualizzare i valori corrispondenti al 25°, 50° e 75° percentile.

Funzioni aggiuntive del comando NPAR TESTS (K campioni dipendenti)

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
28

Analisi a risposta multipla

Sono disponibili due procedure per l'analisi di insiemi a dicotomie e a categorie multiple. La procedura Risposte multiple: Frequenze consente di visualizzare le tabelle di frequenza. La procedura Risposte multiple: Tavole di contingenza consente di visualizzare tavole di contingenza a due o a tre dimensioni. Prima di utilizzare una delle procedure descritte, è necessario definire gli insiemi a risposta multipla.

Esempio. In questo esempio viene illustrato l'utilizzo degli elementi a risposta multipla in un'indagine di mercato. I dati sono fittizi e non devono essere interpretati come reali. È possibile condurre un'indagine tra i passeggeri di una linea aerea in volo su una particolare rotta per ottenere una valutazione della concorrenza. In questo esempio la compagnia American Airlines conduce un'indagine volta a rilevare se i propri passeggeri viaggiano con altre linee aeree sulla rotta Chicago-New York e a determinare l'importanza relativa dei fattori di programmazione e di servizio ai fini della scelta della linea aerea. Al momento dell'imbarco, l'assistente di volo consegna a ciascun passeggero un breve questionario. La prima domanda è la seguente. Tra le seguenti compagnie aeree, contrassegnare quelle utilizzate su questa stessa rotta almeno una volta negli ultimi sei mesi: American, United, TWA, USAir, Altro. Si tratta di una domanda a risposta multipla in quanto il passeggero può indicare più risposte. La domanda, tuttavia, non può essere codificata direttamente in quanto una variabile può contenere un solo valore per ciascun caso. È necessario utilizzare più variabili per associare le risposte a ciascuna domanda. Per eseguire questa operazione è possibile procedere in due modi. Il primo modo consiste nel definire una variabile per ciascuna delle scelte (ad esempio, American, United, TWA, USAir e Altro). Se il passeggero indica la linea United, alla variabile *united* verrà assegnato il codice 1 e in caso contrario il codice 0. Si tratta di un **metodo a dicotomie multiple** per l'assegnazione delle variabili. Il secondo metodo per la classificazione delle risposte è il **metodo a categorie multiple**, che consente di valutare il numero massimo di risposte possibili alla domanda e di impostare lo stesso numero di variabili, con i codici utilizzati per specificare la linea aerea utilizzata. Dall'esame di un campione di questionari può risultare che negli ultimi sei mesi nessun utente ha viaggiato con più di tre diverse linee aeree su questa rotta. Può inoltre risultare che, a causa della deregulation delle linee aree, nella categoria Altro ne vengano citate altre 10. Utilizzando il metodo a risposta multipla, vengono definite tre variabili, ciascuna delle quali è codificata come 1 = *american*, 2 = *united*, 3 = *twa*, 4 = *usair*, 5 = *delta* e così via. Se un determinato passeggero indica American e TWA, alla prima variabile viene assegnato il codice 1, alla seconda il codice 3 e alla terza un codice di valore mancante. Un altro passeggero può aver indicato American e specificato Delta. In questo caso, alla prima variabile viene assegnato il codice 1, alla seconda 5 e alla terza un codice di valore mancante. Se invece si utilizza il metodo a dicotomie multiple, si otterranno 14 variabili distinte. Sebbene ai fini di questa indagine sia possibile utilizzare entrambi i metodi, la scelta del metodo dipende dalla distribuzione delle risposte.

Risposte multiple: Definisci insiemi

La procedura di definizione degli insiemi di variabili a risposta multipla consente di raggruppare le variabili elementari in insiemi a dicotomie o a categorie multiple, per i quali è possibile ottenere tabelle di frequenza e tavole di contingenza. È possibile definire fino a 20 insiemi a risposta multipla. A ogni insieme è necessario assegnare un nome univoco. Per rimuovere un insieme, evidenziarlo nell'elenco dei gruppi a risposta multipla e quindi scegliere Rimuovi. Per modificare un insieme, evidenziarlo nell'elenco, modificare le caratteristiche di definizione desiderate e quindi scegliere Cambia.

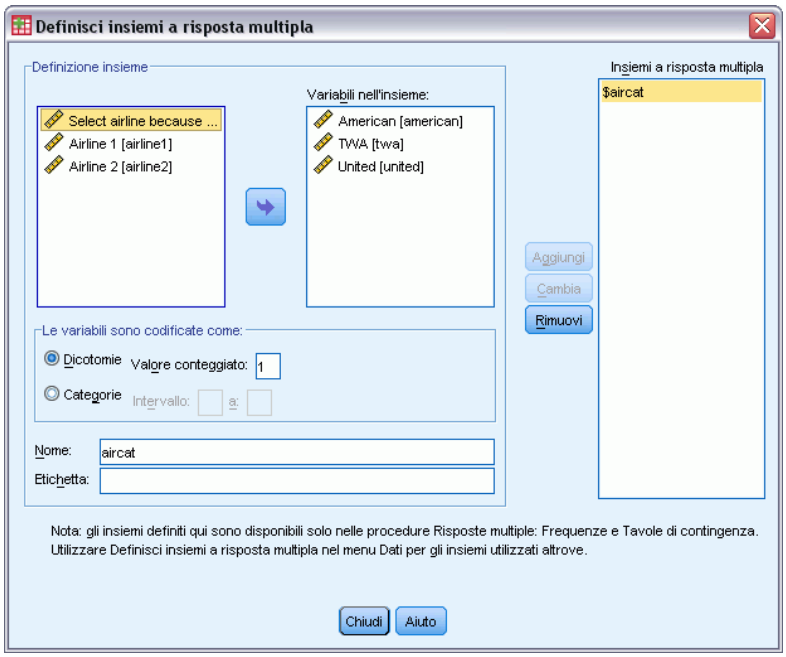
È possibile codificare le variabili elementari come dicotomie o categorie. Per l'utilizzo delle variabili dicotomiche, selezionare Dicotomie per creare un insieme a dicotomie multiple. Specificare un valore intero nella casella Valore conteggiato. Ciascuna variabile contenente almeno un'occorrenza del valore conteggiato diventa una categoria dell'insieme a dicotomie multiple. Selezionare Categorie per creare un insieme a categorie multiple con lo stesso intervallo di valori delle variabili che lo compongono. Specificare valori interi come valori minimo e massimo dell'intervallo di categorie dell'insieme a categorie multiple. Verrà calcolato il totale di ogni singolo valore intero nell'intervallo per tutte le variabili. Le categorie vuote non verranno incluse nella tabella.

A ogni insieme a risposta multipla deve essere assegnato un nome univoco composto al massimo da sette caratteri. Al nome assegnato verrà aggiunto automaticamente il prefisso $ (segno di dollaro). Non è possibile utilizzare i seguenti nomi riservati: *casenum*, *sysmis*, *jdate*, *date*, *time*, *length* e *width*. Il nome dell'insieme a risposta multipla esiste solo ai fini dell'utilizzo in procedure a risposta multipla. Non è possibile fare riferimento ai nomi di insiemi a risposta multipla in altre procedure. È inoltre possibile inserire un'etichetta di variabile descrittiva per l'insieme a risposta multipla. L'etichetta può essere costituita al massimo da 40 caratteri.

Per definire gli insiemi a risposta multipla

- Dai menu, scegliere:
  Analizza > Risposte multiple > Definisci insiemi di variabili...

Figura 28-1
Finestra di dialogo Insiemi a risposta multipla



- Selezionare due o più variabili.
- Se le variabili sono codificate come dicotomie, indicare il valore che si desidera calcolare. Se le variabili sono codificate come categorie, definire l'intervallo delle categorie.
- Immettere un nome univoco per ciascun insieme a risposta multipla.
- Scegliere Aggiungi per aggiungere l'insieme a risposta multipla all'elenco di insiemi definiti.

Risposte multiple: Frequenze

La procedura Risposte multiple: Frequenze consente di ottenere tabelle di frequenza per gli insiemi a risposta multipla. È innanzitutto necessario definire uno o più insiemi a risposta multipla (vedere "Risposte multiple: Definisci insiemi").

Per gli insiemi a dicotomie multiple, i nomi delle categorie indicati nell'output vengono determinati in base alle etichette definite per le variabili elementari del gruppo. Se le etichette delle variabili non sono definite, i nomi delle variabili verranno utilizzati come etichette. Per gli insiemi a categorie multiple, le etichette di categoria vengono determinate in base alle etichette dei valori della prima variabile del gruppo. Se le categorie mancanti per la prima variabile sono presenti per altre variabili del gruppo, definire un'etichetta dei valori per le categorie mancanti.

Valori mancanti. I casi con valori mancanti vengono esclusi tabella per tabella. È inoltre possibile scegliere una delle seguenti opzioni o entrambe:

- **Esclusione listwise all'interno delle dicotomie.**Consente di escludere i casi con valori mancanti per qualsiasi variabile dalla tabella dell'insieme a dicotomie multiple. Questa opzione può essere applicata solo agli insiemi a risposta multipla definiti come insiemi dicotomici. Per

impostazione predefinita, un caso viene considerato mancante per un insieme a dicotomie multiple se nessuna delle variabili che lo compongono contiene il valore conteggiato. I casi con valori mancanti solo per alcune variabili verranno inclusi nelle tabelle del gruppo se almeno una variabile contiene il valore conteggiato.

- **Esclusione listwise all'interno delle categorie.** Consente di escludere i casi con valori mancanti per qualsiasi variabile dalla tabella dell'insieme a categorie multiple. Questa opzione viene applicata solo a insiemi a risposta multipla definiti come insiemi di categorie. Per impostazione predefinita, un caso viene considerato mancante per un insieme a categorie multiple solo se nessuno dei componenti contiene valori validi all'interno dell'intervallo definito.

Esempio. Ogni variabile creata in base a una domanda di un questionario è una variabile elementare. Per analizzare un elemento di un insieme a risposta multipla, è necessario unire le variabili in uno dei due tipi di insiemi a risposta multipla: un insieme a dicotomie multiple oppure un insieme a categorie multiple. Se ad esempio in un'indagine sulle linee aeree dove viene richiesto con quale delle tre linee aeree indicate (American, United, TWA) si è viaggiato negli ultimi sei mesi sono state utilizzate variabili dicotomiche e si è definito un **insieme a dicotomie multiple**, ciascuna delle tre variabili dell'insieme diventerà una categoria della variabile di gruppo. I conteggi e le percentuali relativi alle tre linee aeree verranno visualizzati in una sola tabella di frequenza. Se risulta che nessuna persona ha indicato più di due linee aeree, è possibile creare due variabili, ciascuna con tre codici, ovvero uno per ogni linea aerea. Se si definisce un **insieme a categorie multiple**, i valori verranno inseriti nella tabella aggiungendo gli stessi codici alle variabili elementari. L'insieme di valori risultante equivale agli insiemi di ciascuna variabile elementare. Trenta risposte per United rappresentano ad esempio la somma delle cinque risposte per United per la linea aerea 1 e delle venticinque risposte per United per la linea aerea 2. I conteggi e le percentuali relativi alle tre linee aeree verranno visualizzati in una sola tabella di frequenza.

Statistiche. Tabelle di frequenza in cui vengono visualizzati i conteggi, le percentuali delle risposte, le percentuali dei casi, il numero di casi validi e il numero di casi mancanti.

Dati. Utilizzare gli insiemi a risposta multipla.

Assunzioni. I conteggi e le percentuali forniscono un'utile descrizione dei dati provenienti da qualsiasi distribuzione.

Procedure correlate. La procedura Risposte multiple: Definisci insiemi consente di definire insiemi a risposta multipla.

Per ottenere le frequenze delle risposte multiple

▶ Dai menu, scegliere:
Analizza > Risposte multiple > Frequenze...

Figura 28-2
Finestra di dialogo Risposte multiple: Frequenze

▶ Selezionare uno o più insiemi a risposta multipla.

Risposte multiple: Tavole di contingenza

La procedura Risposte multiple: Tavole di contingenza consente di incrociare insiemi a risposta multipla definiti, variabili elementari o una combinazione di entrambi. È inoltre possibile ottenere le percentuali nelle celle in base a casi o risposte, modificare il trattamento dei valori mancanti oppure accoppiare le variabili di insiemi diversi. È innanzitutto necessario definire uno o più insiemi a risposta multipla (vedere "Per definire gli insiemi a risposta multipla").

Per gli insiemi a dicotomie multiple, i nomi delle categorie indicati nell'output vengono determinati in base alle etichette definite per le variabili elementari del gruppo. Se le etichette delle variabili non sono definite, i nomi delle variabili verranno utilizzati come etichette. Per gli insiemi a categorie multiple, le etichette di categoria vengono determinate in base alle etichette dei valori della prima variabile del gruppo. Se le categorie mancanti per la prima variabile sono presenti per altre variabili del gruppo, definire un'etichetta dei valori per le categorie mancanti. Le etichette delle categorie per le colonne verranno visualizzate su tre righe, composte al massimo da otto caratteri ciascuna. Per evitare di troncare le parole, è possibile invertire le righe e le colonne oppure ridefinire le etichette.

Esempio. In questa procedura è possibile incrociare gli insiemi a dicotomie e a categorie multiple con altre variabili. In un questionario riservato ai passeggeri delle linee aeree venivano richieste le seguenti informazioni: Contrassegnare tra le seguenti tutte le linee aeree di cui ci si è serviti almeno una volta negli ultimi sei mesi (American, United, TWA). Cosa è più importante per la scelta di un volo, la programmazione o il servizio offerto? Scegliere una sola opzione. Dopo aver inserito i dati come dicotomie o categorie multiple e averli uniti in un insieme, è possibile incrociare le domande relative alle linee aeree con quelle relative al servizio o alla programmazione.

Statistiche. Tavole di contingenza con conteggi relativi a celle, righe, colonne e totale e percentuali relative a celle, righe, colonne e totale. Le percentuali nelle celle possono basarsi sui casi o sulle risposte.

Dati. Utilizzare insiemi a risposta multipla oppure variabili categoriali numeriche.

Assunzioni. I conteggi e le percentuali forniscono un'utile descrizione dei dati provenienti da qualsiasi distribuzione.

Procedure correlate. La procedura Risposte multiple: Definisci insiemi consente di definire insiemi a risposta multipla.

Per ottenere tavole di contingenza a risposta multipla

▶ Dai menu, scegliere:
Analizza > Risposte multiple > Tavole di contingenza...

Figura 28-3
Finestra di dialogo Risposte multiple: Tavole di contingenza

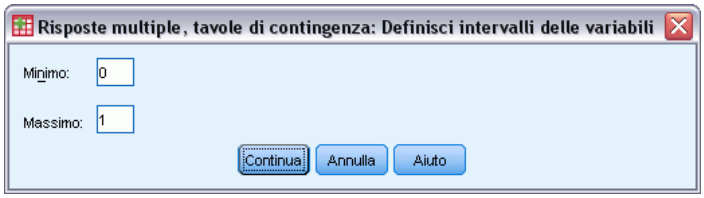
▶ Selezionare una o più variabili numeriche o insiemi a risposta multipla per ciascuna dimensione delle tavole di contingenza.

▶ Definire l'intervallo di ciascuna variabile elementare.

È inoltre possibile ottenere una tavola di contingenza a due vie per ciascuna categoria di una variabile di controllo o di un insieme a risposta multipla. Selezionare uno o più elementi dall'elenco Strati.

Risposte multiple, tavole di contingenza: Definisci intervalli delle variabili

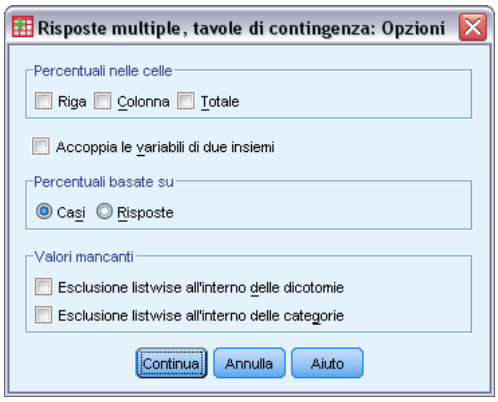
Figura 28-4
Finestra di dialogo Risposte multiple, tavole di contingenza: Definisci intervalli delle variabili



È necessario definire gli intervalli dei valori per tutte le variabili elementari delle tavole di contingenza. Specificare il valore di categoria minimo e massimo che si desidera inserire nelle tavole di contingenza. Le categorie che non rientrano nell'intervallo verranno escluse dall'analisi. Si assume che i valori inclusi nell'intervallo siano interi (i non interi verranno troncati).

Risposte multiple, tavole di contingenza: Opzioni

Figura 28-5
Finestra di dialogo Risposte multiple, tavole di contingenza: Opzioni



Percentuali nelle celle. I conteggi di cella vengono sempre visualizzati. È possibile impostare la visualizzazione di percentuali di riga, percentuali di colonna e percentuali per tabelle a due vie (totale).

Percentuali basate su. È possibile basare le percentuali nelle celle sui casi (o persone che rispondono). Questa opzione non è disponibile se si seleziona la corrispondenza delle variabili tra insiemi a categorie multiple. È inoltre possibile basare le percentuali nelle celle sulle risposte. Per gli insiemi a dicotomie multiple, il numero di risposte equivale al numero di valori conteggiati nei diversi casi. Per gli insiemi a categorie multiple, il numero di risposte equivale al numero di valori dell'intervallo definito.

Valori mancanti. È possibile scegliere una delle seguenti opzioni o entrambe:

- **Esclusione listwise all'interno delle dicotomie.** Consente di escludere i casi con valori mancanti per qualsiasi variabile dalla tabella dell'insieme a dicotomie multiple. Questa opzione può essere applicata solo agli insiemi a risposta multipla definiti come insiemi dicotomici. Per impostazione predefinita, un caso viene considerato mancante per un insieme a dicotomie multiple se nessuna delle variabili che lo compongono contiene il valore conteggiato. I casi con valori mancanti solo per alcune variabili verranno inclusi nelle tabelle del gruppo se almeno una variabile contiene il valore conteggiato.
- **Esclusione listwise all'interno delle categorie.** Consente di escludere i casi con valori mancanti per qualsiasi variabile dalla tabella dell'insieme a categorie multiple. Questa opzione viene applicata solo a insiemi a risposta multipla definiti come insiemi di categorie. Per impostazione predefinita, un caso viene considerato mancante per un insieme a categorie multiple solo se nessuno dei componenti contiene valori validi all'interno dell'intervallo definito.

Per impostazione predefinita, quando si incrociano due insiemi a categorie multiple, ciascuna variabile del primo gruppo verrà incrociata con ciascuna variabile del secondo gruppo e quindi verranno sommati i conteggi relativi a ciascuna cella. Alcune risposte, pertanto, potranno comparire più volte nella stessa tabella. È possibile scegliere la seguente opzione:

Accoppia le variabili di due insiemi. Consente di associare la prima variabile del primo gruppo con la prima variabile del secondo gruppo e così via. Se viene selezionata questa opzione, le percentuali nelle celle saranno basate sulle risposte e non sulle persone che rispondono. Questa opzione non è disponibile per gli insiemi a dicotomie multiple né per le variabili elementari.

Funzioni aggiuntive del comando MULT RESPONSE

Il linguaggio della sintassi dei comandi consente inoltre di:

- Ottenere tavole di contingenza con un massimo di cinque dimensioni (con il sottocomando `BY`).
- Modificare le opzioni di formattazione dell'output, inclusa l'eliminazione delle etichette dei valori (con il sottocomando `FORMAT`).

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
29

Risultati di report

Gli elenchi dei casi e le statistiche descrittive sono strumenti fondamentali per lo studio e la presentazione dei dati. Per creare elenchi dei casi è possibile utilizzare l'Editor dei dati o la procedura Riassumi, per produrre conteggi di frequenze e statistiche descrittive è possibile utilizzare la procedura Frequenze, mentre per creare statistiche per la sottopopolazione è possibile utilizzare la procedura Medie. Queste procedure utilizzano un formato progettato per rendere chiare le informazioni. Per visualizzare le informazioni in un formato diverso, è possibile impostare la presentazione dei dati mediante le opzioni per i report Riepiloghi per righe e Riepiloghi per colonne.

Report : Riepiloghi per righe

La procedura Report: Riepiloghi per righe consente di creare report in cui statistiche riassuntive diverse sono disposte in righe distinte. Sono inoltre disponibili gli elenchi dei casi, che possono includere o meno le statistiche riassuntive.

Esempio. Una ditta proprietaria di una catena di negozi al dettaglio registra le informazioni sui dipendenti, che includono informazioni sugli stipendi e le mansioni nonché sul negozio e il reparto in cui lavora ogni dipendente. È quindi possibile creare un report che includa le informazioni relative a ogni impiegato (elenco) suddivise per negozio e per reparto (variabili di separazione) e che includa statistiche riassuntive (ad esempio, lo stipendio medio) per ciascun negozio e reparto nonché per ciascun reparto di ogni negozio.

Colonne dati. Elenca le variabili da rappresentare nel report, per le quali si desidera creare elenchi dei casi o statistiche riassuntive e controlla il formato per la visualizzazione delle colonne di dati.

Colonne di separazione. Elenca le variabili di separazione facoltative che suddividono il report in più gruppi e consente di gestire le statistiche riassuntive e i formati per la visualizzazione delle colonne di separazione. Se sono presenti più variabili di separazione, per ogni categoria di ciascuna variabile di separazione verrà creato un gruppo distinto all'interno delle categorie della variabile di separazione precedente nell'elenco. Le variabili di separazione devono essere variabili categoriali discrete che suddividono i casi in un numero limitato di categorie significative. I singoli valori di ciascuna variabile di separazione vengono visualizzati, ordinati, in una colonna distinta a sinistra di tutte le colonne dati.

Report. Consente di controllare le caratteristiche generali del report, inclusi i titoli, le statistiche riassuntive globali, la visualizzazione dei valori mancanti e la numerazione delle pagine.

Visualizza casi. Consente di visualizzare i valori effettivi (o etichette dei valori) delle variabili delle colonne di dati per ciascun caso. In tal modo viene creato un listato che potrebbe risultare molto più lungo di un report riepilogativo.

Bozza. Consente di visualizzare solo la prima pagina del report. Questa opzione è utile per visualizzare in anteprima il formato del report prima di generarlo.

I dati sono già ordinati. Se il report include variabili di separazione, prima di generarlo è necessario ordinare il file dati in base ai valori delle variabili di separazione. Se il file di dati è già ordinato in base ai valori delle variabili di separazione, è possibile ridurre i tempi di elaborazione selezionando questa opzione. Questa opzione risulta particolarmente utile dopo aver visualizzato un'anteprima del report.

Per ottenere un riepilogo: Riepiloghi per righe

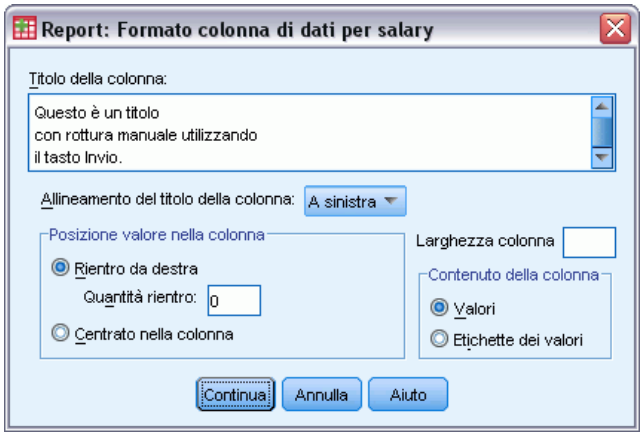
- Dai menu, scegliere:
  Analizza > Report > Report: Riepiloghi per righe...
- Selezionare una o più variabili per Colonne dati. Per ogni variabile selezionata verrà generata una colonna nel report.
- Per i report ordinati e visualizzati in base ai sottogruppi, selezionare una o più variabili per Colonne di separazione.
- Per i report con statistiche riassuntive per i sottogruppi definiti in base alle variabili di separazione, selezionare la variabile di separazione nell'elenco Variabili colonna di separazione e fare clic su Riepilogo nel gruppo Colonne di separazione per specificare le misure riassuntive.
- Per i report con statistiche riassuntive globali, fare clic su Riepilogo per specificare le misure riassuntive.

Figura 29-1
Finestra di dialogo Report: Riepiloghi per righe

Formato delle colonne e di separazione del report

Nelle finestre di dialogo relative al formato è possibile impostare i titoli e la larghezza delle colonne, l'allineamento del testo e la visualizzazione dei valori e delle etichette dei valori. L'opzione Formato colonna dati consente di impostare il formato delle colonne dati nella parte destra della pagina del report. L'opzione Formato di separazione consente di impostare il formato delle colonne di separazione nella parte sinistra.

Figura 29-2
Finestra di dialogo Report: Formato colonna dati



Titolo della colonna. Consente di impostare il titolo della colonna per la variabile selezionata. I titoli lunghi vanno a capo automaticamente all'interno della colonna. Per inserire manualmente le separazioni di riga nella posizione in cui si desidera che i titoli vadano a capo, è possibile utilizzare il tasto Invio.

Posizione valore nella colonna . Per la variabile selezionata, consente di impostare l'allineamento dei valori o delle etichette dei valori all'interno della colonna. L'allineamento dei valori o delle etichette non modifica l'allineamento delle intestazioni di colonna. È possibile rientrare il contenuto delle colonne di un numero specifico di caratteri oppure centrarlo.

Contenuto della colonna. Per la variabile selezionata, consente di impostare la visualizzazione dei valori o delle etichette dei valori definite. I valori per i quali non è stata definita alcuna etichetta vengono sempre visualizzati. Non è disponibile per le colonne dati nei report di riepilogo per colonne.

Report: Linee riassuntive per/Linee riassuntive finali

Le due finestre di dialogo Report: Linee riassuntive consentono di impostare la visualizzazione delle statistiche riassuntive per i gruppi di interruzione e per l'intero report. Linee riassuntive consente di impostare le statistiche dei sottogruppi per ciascuna categoria definita tramite le variabili di separazione. Linee riassuntive finali consente di impostare le statistiche globali visualizzate nella parte finale del report.

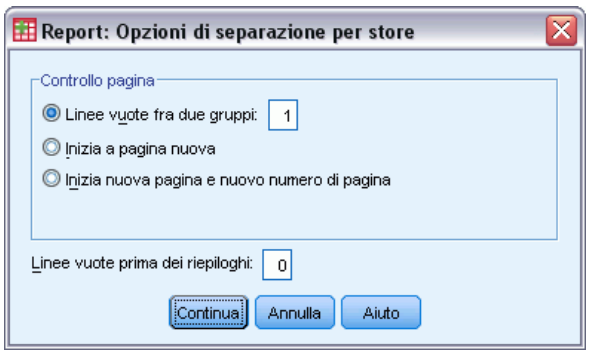
Figura 29-3
Finestra di dialogo Report: Linee riassuntive

Le statistiche riassuntive disponibili sono: somma, media, minimo, massimo, numero di casi, percentuale di casi al di sopra o al di sotto di un valore specifico, percentuale di casi entro un intervallo specifico di valori, deviazione standard, curtosi, varianza e asimmetria.

Report: Opzioni di separazione

La funzione Opzioni di separazione consente di impostare la spaziatura e l'impaginazione delle informazioni sui gruppi.

Figura 29-4
Finestra di dialogo Report: Opzioni di separazione



Controllo pagina. Consente di impostare la spaziatura e l'impaginazione per le categorie relative alla variabile di separazione selezionata. È possibile impostare il numero desiderato di righe vuote tra i gruppi o fare in modo che ciascun gruppo inizi in una nuova pagina.

Linee vuote prima dei riepiloghi. Consente di impostare il numero delle linee vuote tra le etichette dell'asse delle categorie o i dati e le statistiche riassuntive. Questa funzionalità risulta particolarmente utile per i report combinati che includono sia l'elenco dei singoli casi che le statistiche riassuntive per i gruppi. In questo tipo di report è possibile inserire una spaziatura tra l'elenco dei casi e le statistiche riassuntive.

Report: Opzioni

La funzione Report: Opzioni consente di impostare la modalità di elaborazione e la visualizzazione dei valori mancanti e la numerazione delle pagine del report.

Figura 29-5
Finestra di dialogo Report: Opzioni

Esclusione listwise dei valori mancanti. Consente di eliminare dal report i casi con valori mancanti per le variabili del report.

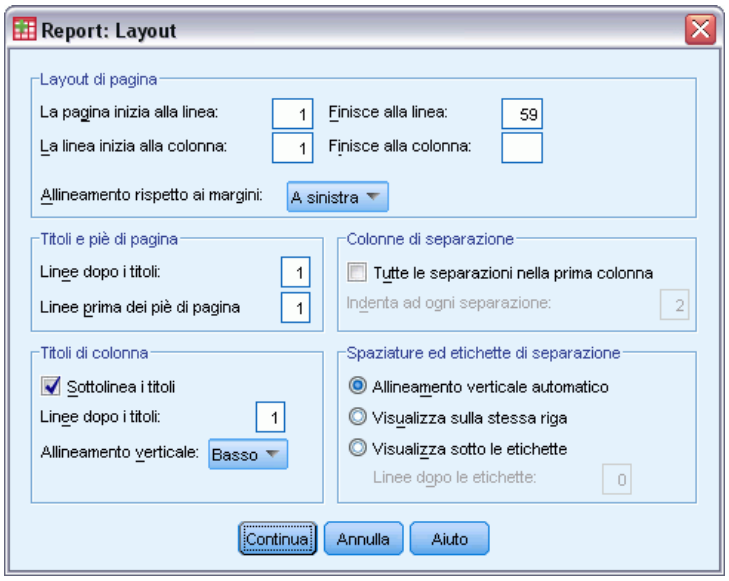
I valori mancanti appaiono come. Consente di specificare il simbolo che rappresenta i valori mancanti nel file dati. Il simbolo deve essere costituito da un solo carattere e viene usato per rappresentare sia i **valori mancanti di sistema** che i **valori mancanti definiti dall'utente**.

Numera pagine a partire da. Consente di specificare un numero di pagina con cui contrassegnare la prima pagina del report.

Report: Layout

L'opzione Report: Layout consente di impostare la larghezza e la lunghezza di ogni pagina del report, la posizione del report nella pagina e l'inserimento di linee vuote ed etichette.

Figura 29-6
Finestra di dialogo Report: Layout



Layout di pagina. Consente di impostare i margini della pagina espressi in linee (superiori e inferiori) e caratteri (a destra e a sinistra) nonché l'allineamento dei report all'interno dei margini.

Titoli e piè di pagina . Consente di impostare il numero delle linee che separano i piè di pagina e i titoli della pagina dal corpo del report.

Colonne di separazione. Consente di impostare la visualizzazione delle colonne di separazione. Se vengono specificate più variabili di separazione, queste possono trovarsi in colonne diverse oppure nella prima colonna. Se tutte le variabili di separazione vengono inserite nella prima colonna, verrà creato un report di larghezza minore.

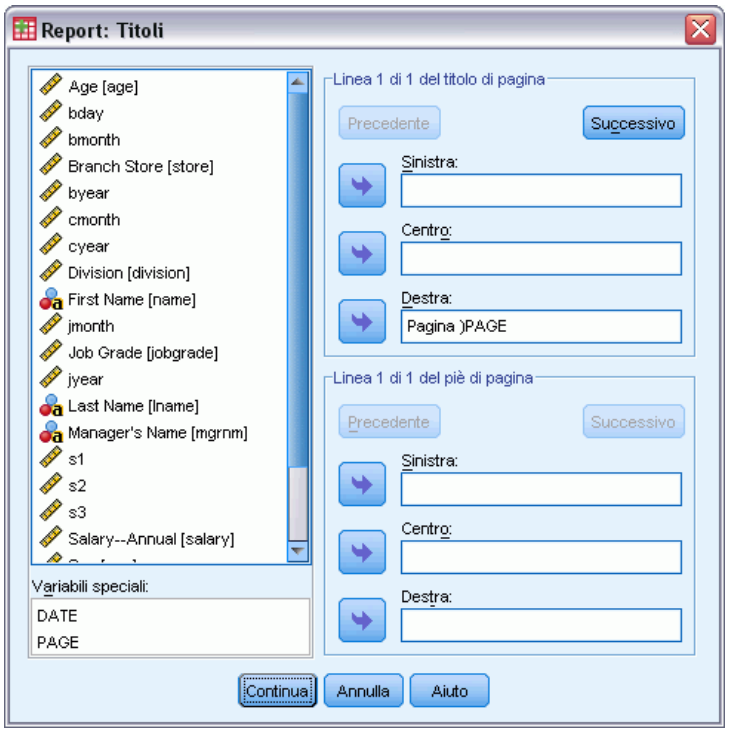
Titoli di colonna. Consente di impostare la visualizzazione dei titoli di colonna, inclusi lo spazio tra i titoli e il corpo del report, la sottolineatura del titolo e l'allineamento verticale dei titoli di colonna.

Spaziature ed etichette di separazione. Consente di posizionare le informazioni relative alle colonne dati (valori e/o statistiche riassuntive) in relazione alle etichette di separazione all'inizio di ogni gruppo. La prima riga delle informazioni sulla colonna dati può iniziare sulla stessa riga in cui si trova l'etichetta di gruppo o dopo un numero specifico di righe rispetto alla posizione dell'etichetta di gruppo. Non è disponibile per i report di riepilogo per colonne.

Report: Titoli

L'opzione Report: Titoli consente di impostare il contenuto e la posizione dei titoli e dei piè di pagina dei report. È possibile specificare fino a dieci righe per i titoli di pagina e per i piè di pagina, con componenti centrati oppure allineati a destra o a sinistra in ciascuna riga.

Figura 29-7
Finestra di dialogo Report: Titoli



Se si inseriscono variabili in titoli e in piè di pagina, l'etichetta dei valori corrente o il valore della variabile verrà visualizzato nel titolo o nel piè di pagina. Nei titoli viene visualizzata l'etichetta dei valori corrispondente al valore della variabile all'inizio della pagina. Nei piè di pagina viene visualizzata l'etichetta dei valori corrispondente al valore della variabile alla fine della pagina. Se non sono presenti etichette dei valori, viene visualizzato il valore effettivo.

Variabili speciali. Le variabili speciali *DATE* e *PAGE* consentono di inserire la data corrente o il numero di pagina in una delle righe dell'intestazione o del piè di pagina del report. Se il file dati utilizzato contiene le variabili denominate *DATE* o *PAGE*, non sarà possibile utilizzare tali variabili nei titoli o nei piè di pagina.

Report: Riepiloghi per colonne

L'opzione Report: Riepiloghi per colonne consente di creare report di riepilogo in cui le statistiche riassuntive vengono visualizzate in colonne distinte.

Esempio. Una ditta proprietaria di una catena di negozi al dettaglio registra le informazioni sui dipendenti. Tra queste sono comprese informazioni sugli stipendi e le mansioni nonché sul reparto in cui lavora ogni dipendente. È quindi possibile creare un report che includa le statistiche riassuntive sugli stipendi (ad esempio media, minimo e massimo) per ogni reparto.

Colonne dati. Consente di visualizzare un elenco delle variabili da rappresentare nel report e per le quali si desidera produrre statistiche riassuntive e di impostare il formato di visualizzazione e le statistiche riassuntive visualizzate per ogni variabile.

Colonne di separazione. Consente di visualizzare l'elenco delle variabili di separazione facoltative che suddividono il report in più gruppi e di impostare i formati di visualizzazione delle colonne di separazione. Se sono presenti più variabili di separazione, per ogni categoria di ciascuna variabile di separazione verrà creato un gruppo distinto all'interno delle categorie della variabile di separazione precedente nell'elenco. Le variabili di separazione devono essere variabili categoriali discrete che suddividono i casi in un numero limitato di categorie significative.

Report. Consente di impostare le caratteristiche globali del report, inclusi i titoli, la visualizzazione dei valori mancanti e la numerazione delle pagine.

Bozza. Consente di visualizzare solo la prima pagina del report. Questa opzione è utile per visualizzare in anteprima il formato del report prima di generarlo.

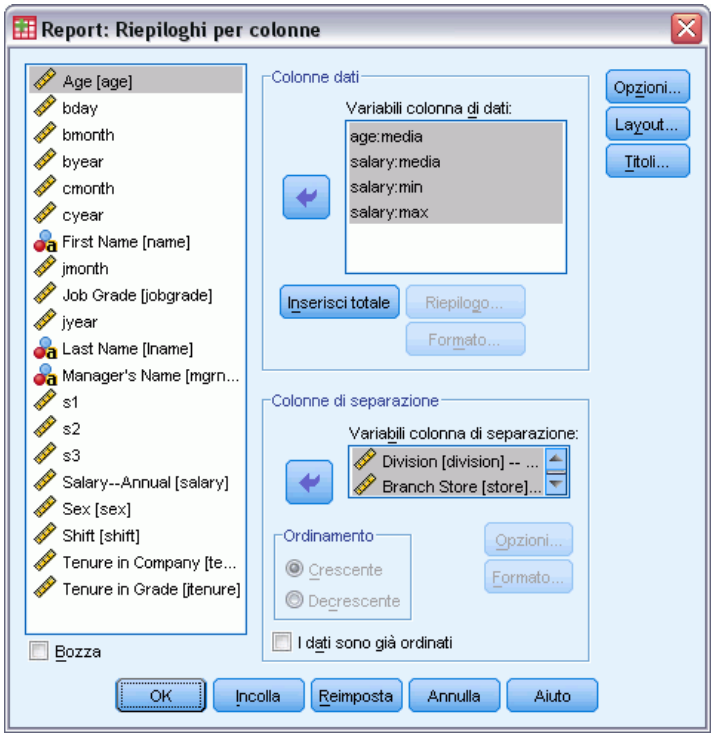
I dati sono già ordinati. Se il report include variabili di separazione, prima di generarlo è necessario ordinare il file dati in base ai valori delle variabili di separazione. Se il file di dati è già ordinato in base ai valori delle variabili di separazione, è possibile ridurre i tempi di elaborazione selezionando questa opzione. Questa opzione risulta particolarmente utile dopo aver visualizzato un'anteprima del report.

Per ottenere un riepilogo: Riepiloghi per colonne

- Dai menu, scegliere:
  Analizza > Report > Report: Riepiloghi per colonne...
- Selezionare una o più variabili per Colonne dati. Per ogni variabile selezionata verrà generata una colonna nel report.

- Per modificare la misura riassuntiva relativa a una variabile, selezionare la variabile nell'elenco Variabili colonna di dati e fare clic su Riepilogo.
- Per ottenere più di una misura riassuntiva per una variabile, selezionare la variabile nell'elenco sorgente e spostarla nell'elenco Variabili colonna di dati più volte, una per ogni misura riassuntiva desiderata.
- Per visualizzare una colonna contenente la somma, la media, il rapporto o altre funzioni per le colonne esistenti, fare clic su Inserisci totale. In tal modo verrà inserita una variabile denominata *totale* nell'elenco Colonne dati.
- Per i report ordinati e visualizzati in base ai sottogruppi, selezionare una o più variabili per Colonne di separazione.

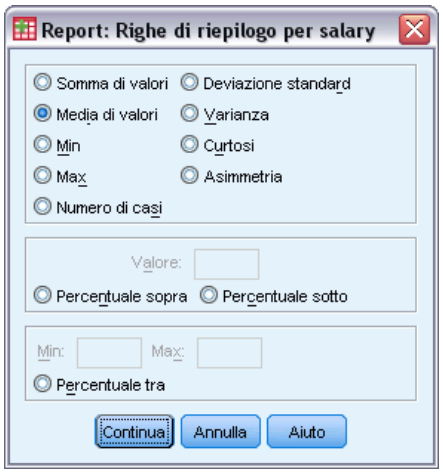
Figura 29-8
Finestra di dialogo Report: Riepiloghi per colonne



Funzione di rappresentazione delle colonne di dati

L'opzione Linee riassuntive consente di controllare le statistiche riassuntive visualizzate per la variabile della colonna dati selezionata.

Figura 29-9
Finestra di dialogo Report: Linee riassuntive



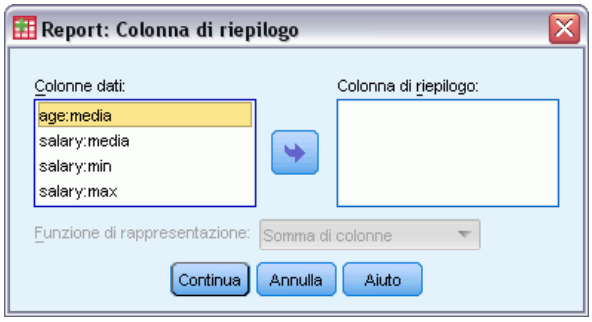
Le statistiche riassuntive disponibili sono: somma, media, minimo, massimo, numero di casi, percentuale di casi al di sopra o al di sotto di un valore specifico, percentuale di casi entro un intervallo specifico di valori, deviazione standard, varianza, curtosi e asimmetria.

Colonna di riepilogo del totale generale

L'opzione Colonna di riepilogo consente di gestire le statistiche riassuntive generali che riassumono due o più colonne dati.

Le statistiche riassuntive generali disponibili sono somma di colonne, media di colonne, minimo, massimo, differenza tra valori in due colonne, quoziente dei valori in una colonna divisi per i valori in un'altra colonna e prodotto di valori di colonne moltiplicati insieme.

Figura 29-10
Finestra di dialogo Report: Colonna di riepilogo



Somma di colonne. La colonna *totale* rappresenta la somma delle colonne nell'elenco Colonna di riepilogo.

Media di colonne. La colonna *totale* rappresenta la media delle colonne nell'elenco Colonna di riepilogo.

Minimo di colonne. La colonna *totale* rappresenta il minimo delle colonne nell'elenco Colonna di riepilogo.

Massimo di colonne. La colonna *totale* rappresenta il massimo delle colonne nell'elenco Colonna di riepilogo.

1a colonna – 2a colonna. La colonna *totale* rappresenta la differenza delle colonne nell'elenco Colonna di riepilogo. L'elenco Colonna di riepilogo deve contenere esattamente due colonne.

1a colonna / 2a colonna. La colonna *totale* rappresenta il quoziente delle colonne nell'elenco Colonna di riepilogo. L'elenco Colonna di riepilogo deve contenere esattamente due colonne.

% 1a colonna / 2a colonna. La colonna *totale* rappresenta la percentuale della prima colonna rispetto alla seconda colonna nell'elenco Colonna di riepilogo. L'elenco Colonna di riepilogo deve contenere esattamente due colonne.

Prodotto di colonne. La colonna *totale* rappresenta il prodotto delle colonne nell'elenco Colonna di riepilogo.

Formato delle colonne del report

Le opzioni di formattazione delle colonne dati e di separazione per i report di riepilogo per colonne sono le stesse descritte per i report di riepilogo per righe.

Report: Opzioni di separazione (Riepiloghi per colonne)

La funzione Opzioni di separazione consente di impostare la visualizzazione, la spaziatura e l'impaginazione per i gruppi.

Figura 29-11
Finestra di dialogo Report: Opzioni di separazione

Totale parziale. Consente di impostare la visualizzazione di totali parziali per i gruppi.

Controllo pagina. Consente di impostare la spaziatura e l'impaginazione per le categorie relative alla variabile di separazione selezionata. È possibile impostare il numero desiderato di righe vuote tra i gruppi o fare in modo che ciascun gruppo inizi in una nuova pagina.

Linee vuote prima del totale parziale. Consente di impostare il numero di righe vuote tra i dati dei gruppi e i totali parziali.

Report: Opzioni (Riepiloghi per colonne)

Le opzioni consentono di impostare la visualizzazione dei totali generali e dei valori mancanti e l'impaginazione dei report di riepilogo per colonne.

Figura 29-12
Finestra di dialogo Report: Opzioni

Totale finale. Consente di visualizzare ed etichettare un totale finale per ogni colonna, visualizzato alla fine della colonna.

Valori mancanti. È possibile escludere i valori mancanti dal report o selezionare un solo carattere per indicare i valori mancanti nel report.

Report: Layout per riepiloghi per colonne

Le opzioni di layout per i report di riepilogo per colonne sono le stesse descritte per i report di riepilogo per righe.

Funzioni aggiuntive del comando REPORT

Il linguaggio della sintassi dei comandi consente inoltre di:

- Visualizzare diverse funzioni di riepilogo nelle colonne di una linea riassuntiva.
- Inserire linee riassuntive nelle colonne dati per le variabili diverse dalle variabili della colonna, o per le varie combinazioni (funzioni composte) di funzioni di rappresentazione.
- Utilizzare Mediana, Moda, Frequenza e Percentuale come funzioni di rappresentazione.
- Controllare in modo più preciso il formato di visualizzazione delle statistiche riassuntive.
- Inserire linee vuote in corrispondenza di vari punti nei report.
- Inserire linee vuote dopo ogni *n* caso nei listati.

A causa della complessità della sintassi `REPORT`, può risultare utile, durante la costruzione di un nuovo report con la sintassi, approssimare il report generato dalle finestre di dialogo, copiare e incollare la sintassi corrispondente e ridefinire tale sintassi in modo da farla corrispondere al report specifico.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
30

Analisi di affidabilità

L'analisi di affidabilità consente di studiare le proprietà delle scale di misurazione e degli elementi che le compongono. La procedura Analisi di affidabilità calcola una serie di misure comunemente utilizzate in relazione all'affidabilità della scala e fornisce inoltre informazioni relative alle relazioni tra singoli elementi della scala. I coefficienti di correlazione tra classi possono essere utilizzati per calcolare le stime di affidabilità.

Esempio. Il questionario misura la soddisfazione del cliente in un modo utile? Utilizzando l'analisi di affidabilità, è possibile determinare il grado di correlazione tra gli elementi del questionario, ottenere un indice globale della ripetibilità oppure la concordanza interna della scala in modo globale. È quindi possibile identificare gli elementi del problema che devono essere esclusi dalla scala.

Statistiche. Descrittive per ogni variabile e per la scala, statistiche riassuntive degli elementi, correlazioni e covarianze tra elementi, stime di affidabilità, tabella ANOVA, coefficienti di correlazione tra classi, $T^{quadrato}$di Hotelling e test di additività di Tukey.

Modelli. Sono disponibili i seguenti modelli di affidabilità:

- **Alfa (Cronbach).** È un modello di concordanza interna, basato sulla media di correlazione fra elementi.
- **Divisione a metà.** Questo modello divide la scala in due parti ed esamina la correlazione tra le parti.
- **Guttman.** Questo modello calcola i limiti inferiori di Guttman per una reale affidabilità.
- **Parallelo.** Questo modello presume che tutti gli elementi abbiano varianze e varianze di errore uguali tra le replicazioni.
- **Parallelo esatto.** Questo modello afferma le ipotesi del modello parallelo e assume inoltre medie uguali degli elementi.

Dati. I dati possono essere dicotomici, ordinali oppure intervalli ma devono essere codificati numericamente.

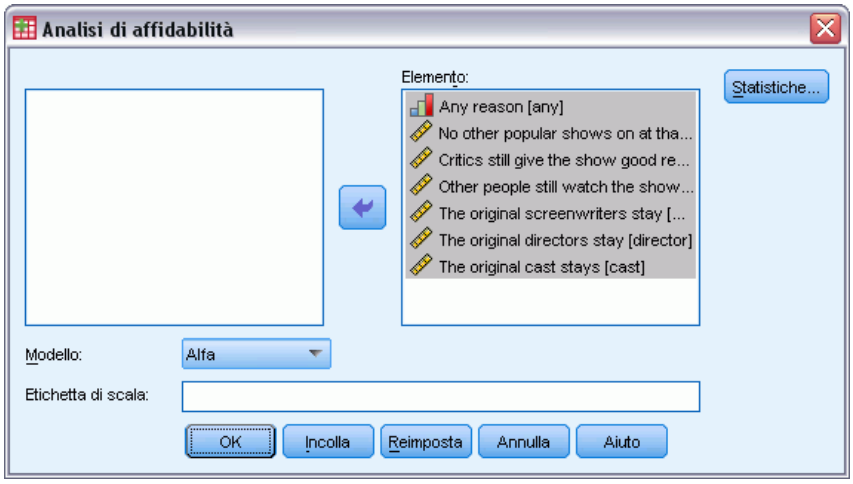
Assunzioni. Le osservazioni devono essere indipendenti e gli errori non devono essere correlati agli elementi. Ogni coppia di elementi deve avere una distribuzione normale bivariata. Le scale devono essere additive, in modo che ogni elemento sia correlato in modo lineare al punteggio totale.

Procedure correlate. Se si desidera esplorare la dimensionalità degli elementi della scala (per verificare se è necessaria più di una costruzione per tenere conto del modello dei punteggi degli elementi), utilizzare la procedura Analisi fattoriale o Scaling multidimensionale. Per identificare i gruppi omogenei di variabili, usare la cluster gerarchica per raggruppare le variabili.

Per ottenere l'analisi di affidabilità

▶ Dai menu, scegliere:
Analizza > Scala > Analisi di affidabilità...

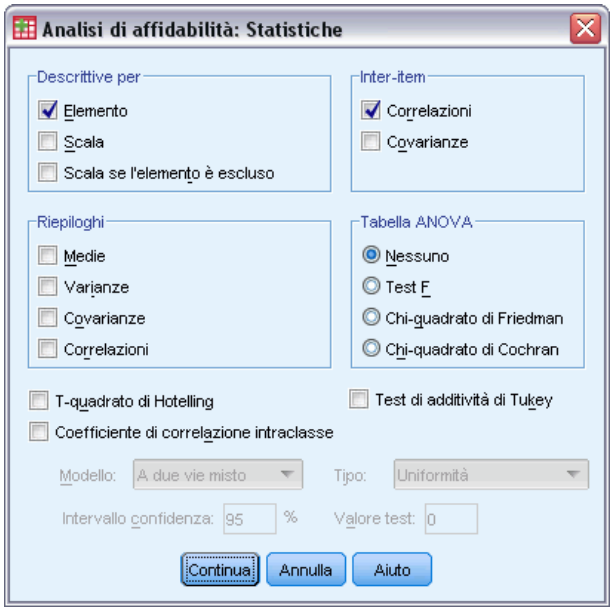
Figura 30-1
Finestra di dialogo Analisi di affidabilità



- Selezionare due o più variabili come potenziali componenti di una scala additiva.
- Scegliere un modello dall'elenco a discesa Modello.

Analisi di affidabilità: Statistiche

Figura 30-2
Finestra di dialogo Analisi di affidabilità: Statistiche



È possibile selezionare varie statistiche per la descrizione della scala e degli elementi. Le statistiche riportate per impostazione predefinita comprendono il numero di casi, il numero di elementi e le stime di affidabilità riportati di seguito:

- **Modelli Alfa.** Per i dati dicotomici, è equivalente al coefficiente Kuder-Richardson 20 (KR20).

- **Modelli Divisione a metà.** Correlazione tra stime dei parametri, affidabilità di Divisione a metà di Guttman, affidabilità di Spearman-Brown (lunghezza uguale e diversa) e coefficiente alfa per ogni metà.
- **Modelli di Guttman.** Coefficienti di affidabilità da lambda 1 a lambda 6.
- **Modelli Parallelo e Parallelo esatto.** Test sulla bontà dell'adattamento del modello, stime della varianza di errore, matrice di correlazione e troncamento, media comune stimata, affidabilità stimata e dati non troncati.

Descrittive per. Fornisce statistiche descrittive per le scale o per gli elementi tra i casi.

- **Elemento.** Fornisce statistiche descrittive per gli elementi tra i casi.
- **Scala.** Fornisce statistiche descrittive per le scale.
- **Scala se l'item è escluso.** Consente di visualizzare statistiche riassuntive per il confronto di ogni elemento con la scala composta dagli altri elementi. Le statistiche includono la media della scala e la varianza risultante se l'elemento venisse eliminato dalla scala, la correlazione tra l'elemento e la scala composta dagli altri elementi e l'Alfa di Cronbach risultante se l'elemento venisse eliminato dalla scala.

Riepiloghi. Fornisce statistiche descrittive della distribuzione di elementi tra tutti gli elementi nella scala.

- **Medie.** Statistiche riassuntive sulle medie degli item. Vengono visualizzate la media più piccola, la più grande e la media centrale, nonché l'intervallo e la deviazione standard delle medie e il rapporto fra la media più grande e la più piccola.
- **Varianze.** Statistiche riassuntive per le varianze degli elementi. Vengono riprodotte la varianza minima, massima e media, nonché l'intervallo e il rapporto fra varianza massima e minima.
- **Covarianze.** Statistiche riassuntive basate sulle correlazioni tra item. Vengono visualizzate la covarianza più piccola, la più grande e la covarianza media, nonché l'intervallo e la deviazione standard delle covarianze e il rapporto fra la covarianza più grande e la più piccola.
- **Correlazioni.** Statistiche riassuntive basate sulle correlazioni tra item. Vengono visualizzate la correlazione più piccola, la più grande e la correlazione media, nonché l'intervallo e la deviazione standard delle correlazioni e il rapporto fra la correlazione più grande e la più piccola.

Inter-item. Fornisce matrici di correlazioni o covarianze tra elementi.

Tabella ANOVA. Fornisce test di medie uguali.

- **Test F.** Visualizza una tabella di analisi della varianza a misure ripetute.
- **Chi-quadrato di Friedman.** Visualizza il chi-quadrato di Friedman e il coefficiente di concordanza di Kendall. Questa opzione è appropriata per dati che rappresentano classifiche (ranghi). Il test chi-quadrato sostituisce il test F solitamente utilizzato nelle tabelle ANOVA.
- **Chi-quadrato di Cochran.** Visualizza la Q di Cochran. Questa opzione è adatta ai dati dicotomici. La Q di Cochran sostituisce il test F solitamente utilizzato nelle tabelle ANOVA.

T quadrato di Hotelling. Crea un test multivariato dell'ipotesi nulla in base alla quale tutti gli elementi sulla scala hanno la stessa media.

Test di additività di Tukey. Crea un test dell'ipotesi in base alla quale non esiste un'interazione moltiplicativa tra gli elementi.

Coefficiente di correlazione intraclasse. Crea misurazioni della consistenza o dell'accordo dei valori all'interno dei casi.

- **Modello.** Consente di selezionare il modello per calcolare i coefficienti di correlazione intraclasse. I modelli disponibili sono A due vie misto, A due vie casuale e A una via casuale. Selezionare A due vie misto quando gli effetti relativi alle persone sono casuali e quelli relativi all'elemento sono fissi, A due vie casuale quando sia gli effetti relativi alle persone che quelli relativi all'elemento sono casuali oppure A una via casuale quando gli effetti relativi alle persone sono casuali.
- **Tipo.** Consente di selezionare il tipo di indice. I tipi disponibili sono Uniformità e Concordanza assoluta.
- **Intervallo di confidenza.** Consente di specificare il livello relativo all'intervallo di confidenza. Il valore predefinito è il 95%.
- **Valore test.** Consente di specificare il valore ipotizzato del coefficiente relativo al test dell'ipotesi. Si tratta del valore rispetto al quale viene confrontato il valore osservato. Il valore predefinito è 0.

Opzioni aggiuntive del comando RELIABILITY

Il linguaggio della sintassi dei comandi consente inoltre di:

- Leggere ed analizzare una matrice di correlazione.
- Scrivere una matrice di correlazione per analizzarla in seguito.
- Specificare le divisioni diverse dalle metà uguali per il metodo di divisione a metà.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
31

Scaling multidimensionale

La procedura Scaling multidimensionale consente di effettuare un tentativo per trovare la struttura in un insieme di misure di distanza tra oggetti o casi. Questa operazione viene compiuta assegnando le osservazioni a posizioni specifiche in uno spazio concettuale (in genere bi o tridimensionale) in modo che le distanze tra i punti nello spazio corrispondano il più possibile alle dissimilarità specificate. In molti casi, le dimensioni di questo spazio concettuale possono essere interpretate ed utilizzate allo scopo di comprendere meglio i dati.

Se si dispone di variabili misurate oggettivamente, è possibile utilizzare lo scaling multidimensionale come una tecnica di riduzione dei dati (la procedura calcolerà le distanze dai dati multivariati per l'utente, se necessario). È inoltre possibile applicare lo scaling multidimensionale alle stime soggettive di dissimilarità tra oggetti o concetti. In aggiunta, Scaling multidimensionale può gestire i dati di dissimilarità da più origini, come nel caso di più stimatori o di rispondenti ai questionari.

Esempio. In che modo i consumatori percepiscono le similitudini tra automobili diverse? Se si dispone di dati che rilevano similarità tra diverse forme e modelli di automobili, lo scaling multidimensionale consentirà di identificare le dimensioni in grado di descrivere le percezioni dei consumatori. È possibile verificare, ad esempio, che il prezzo e le dimensioni di un veicolo definiscono uno spazio bidimensionale, secondo le spiegazioni fornite dai rispondenti.

Statistiche. Per ciascun modello: matrice dei dati, matrice di dati scalati ottimale, s-stress (di Young), stress (di Kruskal), RSQ, coordinate degli stimoli, media dello stress e RSQ per ogni stimolo (modelli RMDS). Per modelli di singole differenze (INDSCAL): pesi di soggetto e valore non ponderato per ogni soggetto. Per ogni matrice in modelli di scaling multidimensionale replicati: stress e RSQ per ogni stimolo. Grafici: coordinate degli stimoli (bi o tridimensionali), grafico a dispersione delle disparità rispetto alle distanze.

Dati. Se si dispone di dati di dissimilarità, tutte le dissimilarità dovrebbero essere quantitative e dovrebbero essere misurate in base alla stessa metrica. Se i dati sono multivariati, le variabili possono essere quantitative, binarie o dati di conteggio. Lo scaling delle variabili è una questione importante poiché le differenze possono influenzare la soluzione. Se le variabili hanno differenze significative (ad esempio, una variabile è misurata in dollari e l'altra è misurata in anni), è consigliabile standardizzarle. Questa operazione può essere eseguita automaticamente dalla procedura Scaling multidimensionale.

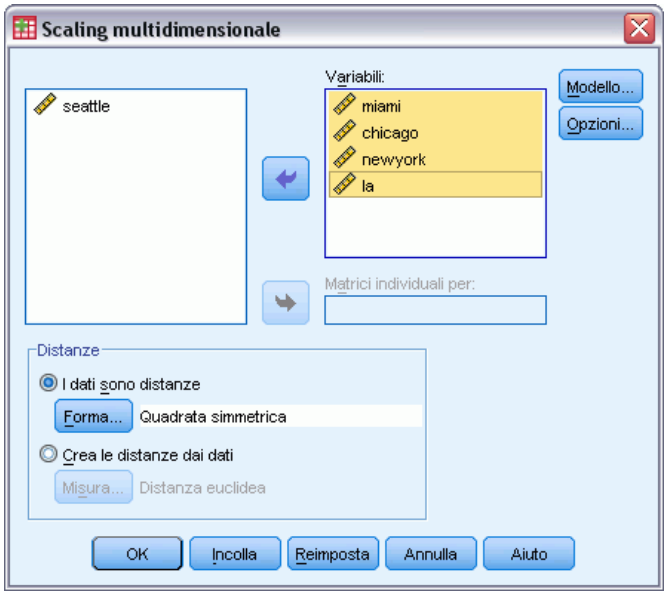
Assunzioni. La procedura Scaling multidimensionale è relativamente libera da ipotesi di distribuzione. Assicurarsi di selezionare il livello di misurazione appropriato (ordinale, intervallo o rapporto) nella finestra di dialogo Scaling multidimensionale per essere sicuri che i risultati vengano calcolati correttamente.

Procedure correlate. Se l'obiettivo è la riduzione dei dati, si può considerare un metodo alternativo quale l'analisi fattoriale, in particolare se le variabili sono di tipo quantitativo. Se si desidera identificare gruppi di casi simili, considerare la possibilità di integrare l'analisi di scaling multidimensionale con un'analisi gerarchica o cluster *k*-medie.

Per ottenere un'analisi Scaling multidimensionale

- Dai menu, scegliere:
  Analizza > Scala > Scaling multidimensionale...

Figura 31-1
Finestra di dialogo Scaling multidimensionale



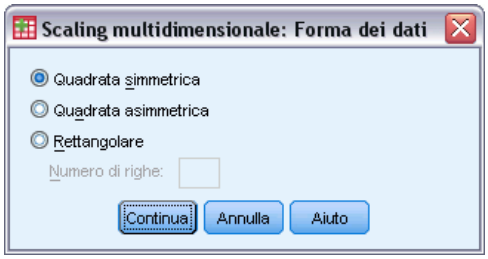
- Selezionare almeno quattro variabili numeriche per l'analisi.
- Nel gruppo Distanze selezionare I dati sono distanze oppureCrea le distanze dai dati.
- Se si seleziona Crea le distanze dai dati, è possibile selezionare anche una variabile di raggruppamento per singole matrici. Le variabili di raggruppamento possono essere numeriche o stringhe.

In alternativa, è possibile anche:

- Se i dati sono distanze, specificare la forma della matrice delle distanze.
- Quando si creano distanze a partire da dati, specificare la misura della distanza da utilizzare.

Scaling multidimensionale: Forma dei dati

Figura 31-2
Finestra di dialogo Forma scaling multidimensionale



Se il file dati attivo rappresenta le distanze tra uno o due insiemi di oggetti, è necessario specificare la forma della matrice dei dati in modo da ottenere i risultati corretti.

Nota: Non è possibile selezionare Quadrata simmetrica se la finestra di dialogo Modello specifica la riga in modo condizionale.

Scaling multidimensionale: Crea misure dai dati

Figura 31-3
Finestra di dialogo Scaling multidimensionale: Crea misure dai dati

La procedura Scaling multidimensionale utilizza dati di dissimilarità per creare una soluzione di scaling. Se i dati disponibili sono dati multivariati (valori di variabili misurate), è necessario creare dati di dissimilarità in modo da calcolare una soluzione di scaling multidimensionale. È possibile specificare i dettagli della creazione delle misure di dissimilarità a partire dai dati disponibili.

Misura. Consente di specificare la misura di dissimilarità per l'analisi. Selezionare un'alternativa dal gruppo Misura corrispondente al tipo di dati desiderato e quindi selezionare una delle misure dall'elenco a discesa corrispondente a tale tipo di misura. Le alternative disponibili sono:

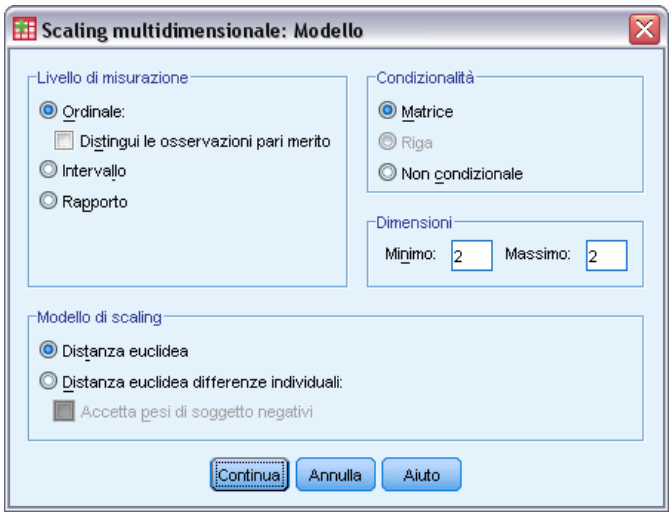
- **Intervallo**. Distanza euclidea, Distanza euclidea quadratica, Chebychev, City-Block, Minkowski o Personalizzato.
- **Conteggi**. Misura chi-quadrato e Misura phi-quadrato.
- **Binaria**. Distanza euclidea, Distanza euclidea quadratica, Differenza di dimensione, Differenza di modello, Varianza o Lance e Williams.

Crea matrice delle distanze. Consente di scegliere l'unità di analisi. Le alternative sono Fra variabili o Fra casi.

Trasforma valori. In alcuni casi, ad esempio quando le variabili sono misurate su scale molto diverse, è possibile standardizzarne i valori prima di calcolare le dissimilarità (non applicabile ai dati binari). Selezionare un metodo di standardizzazione dall'elenco a discesa Standardizza. Se non è richiesta alcuna standardizzazione, selezionare Nessuna.

Scaling multidimensionale: Modello

Figura 31-4
Finestra di dialogo Scaling multidimensionale: Modello



La stima corretta di un modello di scaling multidimensionale dipende dagli aspetti dei dati e del modello stesso.

Livello di misurazione. Consente di specificare il livello dei dati. Le alternative sono Ordinale, Intervallo o Rapporto. Se le variabili sono ordinali, la selezione dell'opzione Distingui le osservazioni pari merito richiede che vengano trattate come variabili continue, in modo che i pari merito (valori uguali per casi differenti) siano risolti in modo ottimale.

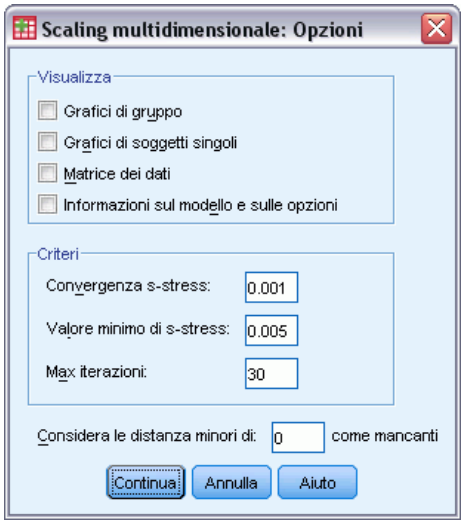
Condizionalità. Consente di specificare quali confronti sono significativi. Le alternative sono Matrice, Riga o Non condizionale.

Dimensioni. Consente di specificare la dimensione della soluzione di scaling. Per ciascun numero nell'intervallo viene calcolata una soluzione. Specificare interi da 1 a 6. È consentito un minimo di 1 solo se si è selezionato Distanza euclidea come modello di scaling. Per una soluzione singola, specificare lo stesso valore minimo e massimo.

Modello di scaling. Consente di specificare le ipotesi in base alle quali viene eseguito lo scaling. Le alternative disponibili sono Distanza euclidea o Distanza euclidea per differenze singole (nota anche come INDSCAL). Nel modello Distanza euclidea per differenze singole, è possibile selezionare l'opzione Accetta pesi di soggetto negativi, se appropriata per i dati disponibili.

Scaling multidimensionale: Opzioni

Figura 31-5
Finestra di dialogo Scaling multidimensionale: Opzioni



È possibile impostare le opzioni per l'analisi scaling multidimensionale:

Visualizzazione. Consente di selezionare vari tipi di output. Le opzioni disponibili sono Grafici di gruppo, Grafici di soggetti singoli, Matrice dei dati e Informazioni sul modello e sulle opzioni.

Criteri. Consente di determinare il momento in cui interrompere l'iterazione. Per modificare i valori predefiniti, inserire i valori per Convergenza s-stress, Valore minimo di s-stress e Massimo numero di iterazioni.

Considera le distanze minori di N come mancanti. Le distanze minori di tale valore sono escluse dall'analisi.

Opzioni aggiuntive del comando ALSCAL

Il linguaggio della sintassi dei comandi consente inoltre di:

- Utilizzare tre tipi di modelli aggiuntivi, noti come ASCAL, AINDS e GEMSCAL, per lo scaling multidimensionale.
- Eseguire trasformazioni polinomiali su dati di intervallo e di rapporto.
- Analizzare le similarità (piuttosto che le distanze) con dati ordinali.
- Analizzare i dati nominali.
- Salvare varie matrici di coordinate e pesi nei file e rileggerli per l'analisi.
- Vincolare l'unfolding multidimensionale.

Per informazioni dettagliate sulla sintassi, vedere *Command Syntax Reference*.

Capitolo
32

Statistiche di rapporto

La procedura Statistiche di rapporto offre un elenco completo di statistiche riassuntive per la descrizione del rapporto tra due variabili di scala.

È possibile ordinare l'output in base ai valori di una variabile di raggruppamento in ordine crescente o decrescente. È possibile eliminare il report sulle statistiche di rapporto dall'output e salvare i risultati in un file esterno.

Esempio. Esiste un buon grado di uniformità nel rapporto tra prezzo di stima e prezzo di vendita delle case in ognuno dei cinque paesi? Dall'output, si apprende che la distribuzione dei rapporti varia notevolmente da paese a paese.

Statistiche. Mediana, media, media pesata, intervalli di confidenza, coefficiente di dispersione, coefficiente di variazione centrato sulla mediana, coefficiente di variazione centrato sulla media, differenziale di prezzo, deviazione standard, deviazione assoluta media, intervallo, valori minimo e massimo e indice di concentrazione calcolato per una percentuale o un intervallo specifico per l'utente compresi nel rapporto della mediana.

Dati. Utilizzare codici numerici o stringhe per codificare le variabili di raggruppamento (misure di livello nominale o ordinale).

Assunzioni. Le variabili che definiscono il numeratore e il denominatore del rapporto dovrebbero essere variabili di scala con valori positivi.

Per ottenere statistiche di rapporto

▶ Dai menu, scegliere:
Analizza > Statistiche descrittive > Rapporto...

Figura 32-1
Finestra di dialogo Statistiche di rapporto

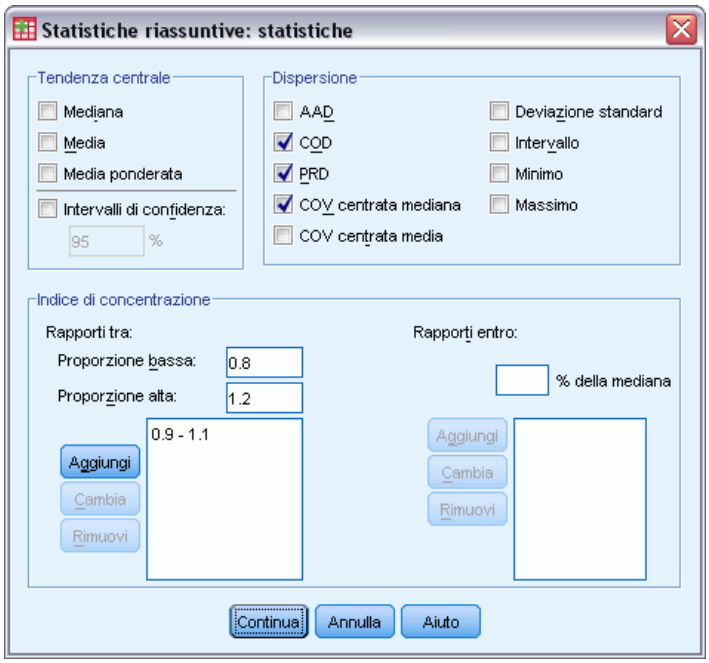
- ▶ Selezionare una variabile numeratore.
- ▶ Selezionare una variabile denominatore.

  Oppure:

  - Selezionare una variabile di raggruppamento e specificare l'ordinamento dei gruppi nei risultati.
  - Scegliere se visualizzare o meno i risultati nel Viewer.
  - Decidere se salvare o meno i risultati in un file esterno per futuri utilizzi. In caso affermativo, specificare il nome del file in cui verranno salvati.

Statistiche di rapporto

Figura 32-2
Finestra di dialogo Statistiche di rapporto



Tendenza centrale. Le misure di tendenza centrale sono statistiche che descrivono la distribuzione dei rapporti.

- **Mediana.** Il valore tale che il numero di rapporti inferiore a questo valore e il numero di rapporti maggiore di questo valore siano uguali.
- **Media.** Il risultato della somma dei rapporti e la divisione del risultato per il numero totale di rapporti.
- **Media ponderata.** Il risultato della divisione della media del numeratore per la media del denominatore. La media pesata è anche la media dei rapporti ponderati dal denominatore.
- **Intervalli di confidenza.** Vengono visualizzati gli intervalli di confidenza per la media, la mediana e la media ponderata (se necessario). Specificare un valore maggiore o uguale a 0 e minore di 100 come intervallo di confidenza.

Dispersione. Queste statistiche misurano il grado di variazione o variabilità nei valori osservati.

- **AAD.** (Deviazione assoluta media) È ottenuta sommando le deviazioni assolute dei rapporti dalla mediana e dividendo il risultato per il numero totale di rapporti.
- **COD.** (Coefficiente di dispersione) È il risultato dell'espressione della deviazione assoluta media come una percentuale della mediana.
- **PRD.** (Differenziale di prezzo) Anche noto come indice di regressività, è il risultato della divisione della media per la media pesata.
- **COV centrata mediana.** (Coefficiente di variazione) È il risultato dell'espressione delle radici quadrate medie della deviazione dalla mediana come una percentuale della mediana.
- **COV centrata media.** (Coefficiente di variazione) È il risultato dell'espressione della deviazione standard come una percentuale della media.
- **Deviazione standard.** La deviazione standard viene ottenuta sommando le deviazioni quadrate dei rapporti dalla media, dividendo il risultato per il numero totale di rapporti meno uno ed estraendo la radice quadrata positiva.
- **Intervallo.** L'intervallo è il risultato della differenza tra rapporto massimo e rapporto minimo.
- **Minimo.** Il minimo è il rapporto più piccolo.
- **Massimo.** Il massimo è il rapporto più grande.

Indice di concentrazione. Consente di misurare la percentuale di rapporti che cade in un intervallo. È possibile calcolarlo in due modi diversi:

- **Rapporti tra.** In questo caso l'intervallo viene definito in modo esplicito specificando i valori alti e bassi dell'intervallo. Immettere i valori per le proporzioni alte e basse, quindi fare clic su Aggiungi per ottenere un intervallo.
- **Rapporti entro.** In questo caso l'intervallo viene definito in modo implicito specificando la percentuale della mediana. Digitare un valore compreso tra 0 e 100 e fare clic su Aggiungi. Il valore minimo dell'intervallo è uguale a (1 – 0,01 × valore) × mediana, e il massimo è uguale a (1 + 0,01 × valore) × mediana.

Capitolo
33

Curve ROC

Questa procedura è utile per valutare le prestazioni degli schemi di classificazione in cui i soggetti sono classificati in base a una variabile con due categorie.

Esempio. Poiché è interesse della banca classificare in modo corretto i clienti adempienti o meno, è necessario elaborare metodi specifici per decidere a chi concedere i prestiti. Le curve ROC possono essere utilizzate per valutare il livello di attendibilità di questi metodi.

Statistiche. Area sotto alla curva ROC con intervallo di confidenza e coordinate della curva ROC. Grafici: Curva ROC.

Metodi. È possibile calcolare una stima dell'area sotto alla curva ROC in modo non parametrico o parametrico mediante un modello esponenziale binegativo.

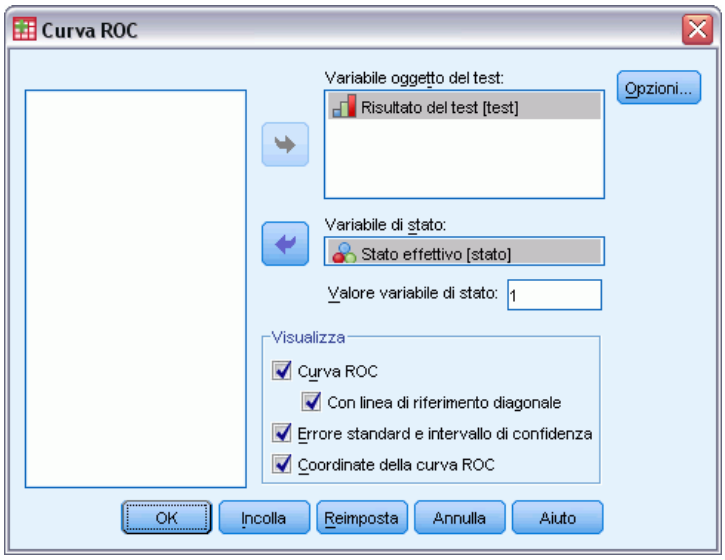
Dati. Le variabili del test sono quantitative. Le variabili del test sono spesso composte dalle probabilità dell'analisi discriminante o della regressione logistica oppure da punteggi su scala arbitraria, i quali rappresentano la "forza della convinzione" di chi classifica i soggetti in una categoria piuttosto che in un'altra. La variabile di stato può essere di qualsiasi tipo e indica la vera categoria a cui appartiene un soggetto. Il valore della variabile di stato indica quale categoria debba essere considerata *positiva*.

Assunzioni. Si suppone che i numeri crescenti sulla scala della classifica rappresentino la convinzione crescente che il soggetto appartenga a una categoria, mentre i numeri decrescenti rappresentino la convinzione crescente che il soggetto appartenga all'altra categoria. L'utente deve scegliere qual è la direzione *positiva*. Si suppone inoltre che la categoria *vero* alla quale appartiene ciascun soggetto sia conosciuta.

Per creare una Curva ROC

- Dai menu, scegliere:
  Analizza > Curva ROC...

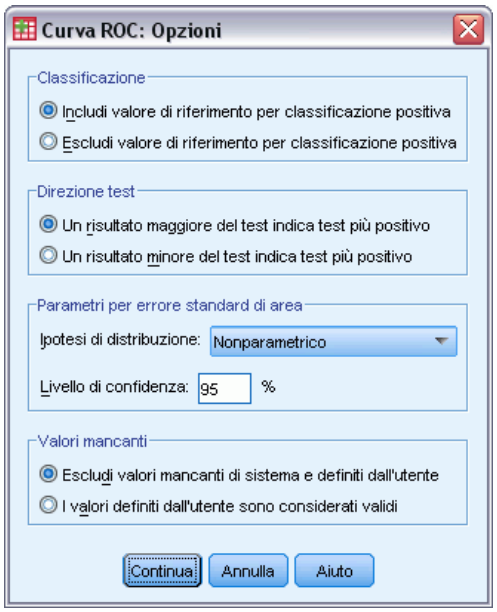
Figura 33-1
Finestra di dialogo Curva ROC



- Selezionare una o più variabili di probabilità per il test.
- Selezionare una variabile di stato.
- Individuare il valore *positivo* per la variabile di stato.

Curva ROC: Opzioni

Figura 33-2
Finestra di dialogo Curva ROC: Opzioni



È possibile selezionare le seguenti opzioni per l'analisi ROC:

Classificazione.. Consente di specificare se il valore di riferimento debba essere incluso o escluso dalla classificazione *positiva*. Questa impostazione non ha attualmente alcun effetto sull'output.

Direzione test. Consente di specificare la direzione della scala in rapporto alla categoria *positiva* .

Parametri per errore standard di area. Consente di specificare il metodo di valutazione dell'errore standard dell'area sotto alla curva. I metodi disponibili sono non parametrico ed esponenziale binegativo. Consente inoltre di impostare il livello dell'intervallo di confidenza. L'intervallo disponibile è da 50,1% a 99,9%.

Valori mancanti. Consente di specificare le modalità di gestione dei valori mancanti.

Appendice

A

Note legali

Queste informazioni sono state preparate per prodotti e servizi offerti in tutto il mondo.

IBM potrebbe non offrire i prodotti, i servizi o le funzionalità di cui si tratta nel presente documento in altri paesi. Contattare il rappresentante IBM locale per informazioni sui prodotti e i servizi attualmente disponibili nella propria zona. Qualsiasi riferimento a un prodotto, programma o servizio IBM non intende dichiarare o implicare che sia possibile utilizzare esclusivamente tale prodotto, programma o servizio IBM. Potrà invece essere utilizzato qualsiasi prodotto, programma o servizio con funzionalità equivalente e che non violi i diritti di proprietà intellettuale di IBM. Tuttavia, è responsabilità dell'utente valutare e verificare il funzionamento di qualsiasi prodotto, programma o servizio non IBM.

IBM può essere titolare di brevetti o domande di brevetto relativi alla materia oggetto del presente documento. La consegna del presente documento non conferisce alcuna licenza rispetto a questi brevetti. Rivolgere per iscritto i quesiti sulle licenze a:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.

Per richieste di informazioni sulle licenze riguardanti il set di caratteri a byte doppio (DBCS), contattare l'Intellectual Property Department di IBM del proprio paese, oppure inviare le richieste in forma scritta all'indirizzo:

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Giappone.

Il seguente paragrafo non si applica per il Regno Unito o altri paesi in cui le presenti disposizioni non sono conformi alle leggi locali: INTERNATIONAL BUSINESS MACHINES FORNISCE QUESTA PUBBLICAZIONE "COSÌ COM'È" SENZA GARANZIA DI ALCUN TIPO, SIA ESSA ESPRESSA O IMPLICITA, INCLUSE, MA NON LIMITATE A, LE GARANZIE IMPLICITE DI NON VIOLAZIONE, COMMERCIABILITÀ O IDONEITÀ A UNO SCOPO SPECIFICO. Alcuni stati non consentono limitazioni di garanzie espresse o implicite in determinate transazioni, pertanto quanto sopra potrebbe non essere applicabile.

Le presenti informazioni possono includere imprecisioni tecniche o errori tipografici. Le modifiche periodiche apportate alle informazioni contenute in questa pubblicazione verranno inserite nelle nuove edizioni della pubblicazione. IBM può apportare miglioramenti e/o modifiche al/ai prodotto/i e/o al/ai programma/i descritti nella presente pubblicazione in qualsiasi momento senza preavviso.

Qualsiasi riferimento nelle presenti informazioni a siti Web non IBM viene fornito esclusivamente per facilitare la consultazione e non rappresenta in alcun modo un'approvazione o sostegno da parte nostra di tali siti Web. I materiali contenuti in tali siti Web non fanno parte dei materiali di questo prodotto IBM e il loro utilizzo è esclusivamente a rischio dell'utente.

IBM può utilizzare o distribuire eventuali informazioni fornite dall'utente nei modi che ritiene appropriati senza incorrere in alcun obbligo nei confronti dell'utente.

I licenziatari del programma che desiderassero informazioni su di esso allo scopo di abilitare: (i) lo scambio di informazioni tra programmi creati indipendentemente e altri programmi (questo compreso) e (ii) l'utilizzo in comune delle informazioni scambiate, dovranno rivolgersi a:

IBM Software Group, All'attenzione di: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

Tali informazioni saranno fornite in conformità ai termini e alle condizioni in vigore e, in alcuni casi, dietro pagamento.

Il programma concesso in licenza descritto nel presente documento e tutto il materiale correlato disponibile sono forniti da IBM in base ai termini del contratto di licenza cliente IBM, del contratto di licenza internazionale IBM o del contratto equivalente esistente tra le parti.

le informazioni relative a prodotti non IBM sono state ottenute dai fornitori di tali prodotti, da loro annunci pubblicati e da altre fonti disponibili al pubblico. IBM non ha verificato tali prodotti e non può confermare l'accuratezza delle prestazioni, la compatibilità o qualsiasi altra dichiarazione relativa a prodotti non IBM. Eventuali domande in merito alle funzionalità dei prodotti non IBM vanno indirizzate ai fornitori di tali prodotti.

Le presenti informazioni includono esempi di dati e report utilizzati in operazioni aziendali quotidiane. Per fornire una descrizione il più possibile esaustiva, gli esempi includono nomi di persone, società, marchi e prodotti. Tutti questi nomi sono fittizi e ogni somiglianza a nomi e indirizzi utilizzati da aziende reali è puramente casuale.

Per chi visualizza queste informazioni a video: le fotografie e le illustrazioni a colori potrebbero non essere disponibili.

Marchi commerciali

IBM, il logo IBM, ibm.com e SPSS sono marchi di IBM Corporation, registrati in numerose giurisdizioni nel mondo. Un elenco aggiornato dei marchi IBM è disponibile sul Web all'indirizzo *http://www.ibm.com/legal/copytrade.shtml*.

Adobe, il logo Adobe, PostScript e il logo PostScript sono marchi o marchi registrati di Adobe Systems Incorporated negli Stati Uniti e/o in altri paesi.

Intel, il logo Intel, Intel Inside, il logo Intel Inside, Intel Centrino, il logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium sono marchi o marchi registrati di Intel Corporation o delle sue consociate negli Stati Uniti e in altri paesi.

Java e tutti i marchi e i logo basati su Java sono marchi di Sun Microsystems, Inc. negli Stati Uniti e/o negli altri paesi.

Linux è un marchio registrato di Linus Torvalds negli Stati Uniti e/o negli altri paesi.

Microsoft, Windows, Windows NT e il logo Windows sono marchi di Microsoft Corporation negli Stati Uniti e/o negli altri paesi.

UNIX è un marchio registrato di The Open Group negli Stati Uniti e in altri paesi.

Questo prodotto utilizza WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Altri nomi di prodotti e servizi possono essere marchi commerciali di IBM o di altre aziende.

Le schermate dei prodotti Adobe sono state ristampate su autorizzazione di Adobe Systems Incorporated.

Le schermate dei prodotti Microsoft sono state ristampate su autorizzazione di Microsoft Corporation.

Indice