IBM SPSS Statistics Base 22

Nota
Prima di utilizzare queste informazioni e il prodotto che supportano, leggere le informazioni in “Avvisi” a pagina 197.

Informazioni sul prodotto

Questa edizione si applica alla versione 22, release 0, modifica 0 di IBM SPSS Statistics e a tutte le release e modifiche successive se non diversamente indicato in nuove edizioni.

Indice

Capitolo 1. Informazioni sui dati

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

Nota: le informazioni sui dati ignorano lo stato del file di suddivisione. Sono inclusi i gruppi di file di suddivisione creati per i valori mancanti valori minimo, massimo e di incremento (disponibili nell'opzione modulo aggiuntivo Valori mancanti).

Per ottenere le informazioni sui dati

1. Dai menu, scegliere:
   Analizza > **Report** > **Informazioni sui dati**
2. Fare clic sulla scheda Variabili.
3. 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 di riepilogo).
- 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 di riepilogo visualizzate. Per ulteriori informazioni, consultare l'argomento “Scheda Informazioni sui dati - Statistiche” a pagina 3.

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)

1. Fare clic con il pulsante destro del mouse su una variabile nell'elenco origine.
2. Selezionare un livello di misurazione dal menu a comparsa.

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.

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 di origine 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 che vengano forniti i primi dati quando il livello di misurazione non è stato impostato in modo esplicito, ad esempio, i dati letti da un'origine esterna o le variabili appena create. Per ulteriori informazioni, consultare l'argomento .

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

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

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

Valori mancanti. Valori mancanti definiti dall'utente. Se si seleziona Conteggio o Percentuale nella scheda Statistiche, le etichette valore definite vengono incluse nell'output anche se non si seleziona Valori mancanti qui. Non è disponibile per gli insiemi a risposta multipla.

Attributi personalizzati. Attributi della variabile definiti dall'utente. L'output include sia i nomi sia i valori di tutti gli attributi della variabile personalizzati associati a ciascuna variabile. Per ulteriori informazioni, consultare l'argomento . Non è disponibile per gli insiemi a risposta multipla.

Attributi riservati. Attributi della variabile di sistema riservati. 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 il dataset non è mai stato salvato in formato IBM 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, il dataset attivo non ha un nome di file.

Ubicazione. Ubicazione della directory (cartella) del file di dati di IBM SPSS Statistics. Se il dataset non è mai stato salvato in formato IBM SPSS Statistics, non esiste alcuna ubicazione.

Numero di casi. Numero di casi presenti nel dataset attivo. Si tratta del numero totale di casi, compresi gli eventuali casi esclusi dalle statistiche di riepilogo 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 del peso. Se il peso è attivo, viene visualizzato il nome della variabile peso. Per ulteriori informazioni, consultare l'argomento .

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 nel dataset (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. Crea 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 che vengano forniti i primi dati quando il livello di misurazione non è stato impostato in modo esplicito, ad esempio, i dati letti da un'origine esterna o le variabili appena create.

Elenco di 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 del criterio di ordinamento include anche i nomi degli attributi della variabile personalizzati definiti 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 valore, conteggi o percentuali per ogni valore univoco, è possibile eliminare queste informazioni dalla tabella se il numero di valori supera il valore specificato. Per impostazione predefinita, queste informazioni non vengono inserite se il numero di valori univoci per la variabile supera 200.

Scheda Informazioni sui dati - Statistiche

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

Conteggi e percentuali

Per le variabili nominali e ordinali, gli insiemi a risposta multipla e i valori con etichetta delle variabili di scala, sono disponibili le statistiche riportate di seguito.

Conteggio. Il conteggio o il numero di casi che hanno ciascun valore (o intervallo di valori) di una variabile.

Percentuale. La percentuale di casi con uno specifico valore.

Tendenza centrale e dispersione

Per le variabili di scala, le statistiche disponibili sono:

Media. Una misura di tendenza centrale. La media aritmetica, ossia la somma divisa per il numero di casi.

Deviazione standard. Una misura della dispersione intorno alla media. 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, l'età media è 45 e la deviazione standard è 10, il 95% dei casi sarà compreso tra 25 e 65 in un distribuzione normale.

Quartili. Visualizza 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 di riepilogo visualizzate per tale variabile) nell'elenco di variabili di origine nella 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 report 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 report sulle frequenze può essere eliminato se una variabile ha molti valori distinti. È possibile etichettare i grafici con frequenze (valore predefinito) 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 della frequenza, percentuali, percentuali cumulative, 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.

Considerazioni sui dati relativi alle frequenze

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

Ipotesi. Le tabulazioni 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 di riepilogo, 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 delle frequenze

1. Dai menu, scegliere:
   Analizza > **Statistiche descrittive** > **Frequenze...**
2. 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

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 avere un numero uguale di gruppi diverso da quattro, selezionare **Punti di divisione per: n 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 l'ubicazione della distribuzione includono media, mediana, moda e somma di tutti i valori.

- *Media*. Una misura di tendenza centrale. La media aritmetica, ossia la somma divisa per il numero di casi.
- *Mediana*. Il valore sopra il quale e sotto il quale ricade la metà dei casi, il 50° 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 estremamente bassi o alti.
- *Modalità*. Il valore che ricorre più spesso. Se più valori condividono la maggiore ricorrenza, ognuno di essi è una moda. La procedura Frequenze riporta solo la più piccola di queste modalità multiple.
- *Somma*. La somma o il totale dei valori, su tutti i casi con valori non mancanti.

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

- *Deviazione std.*. Una misura della dispersione intorno alla media. 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, l'età media è 45 e la deviazione standard è 10, il 95% dei casi sarà compreso tra 25 e 65 in un distribuzione normale.
- *Varianza*. Una misura della dispersione dei valori intorno alla media, pari alla somma dei quadrati delle deviazioni dalla media, divisa per un valore corrispondente al numero totale dei casi meno uno. La varianza è misurata in unità pari al quadrato di quelle della variabile stessa.
- *Intervallo*. La differenza tra il valore massimo ed il valore minimo di una variabile numerica, il massimo meno il minimo.
- *Minimo*. Il valore più basso di una variabile numerica.
- *Massimo*. Il valore più alto di una variabile numerica.
- *E. S. della media*. Una misura di quanto il valore della media può variare da campione a campione per campioni presi dalla stessa distribuzione. Può essere usata per comparare 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 asimmetria positiva ha una 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 uno scostamento dalla simmetria.
- *Curtosi*. Una misura di quanto le osservazioni si raggruppino attorno a un punto centrale. 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

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 diffusione 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 della scala può essere etichettato dalle percentuali o dai conteggi delle frequenze.

Frequenze: Formato

Ordina per. La tabella delle frequenze 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 delle tabelle di statistiche per più variabili, è possibile visualizzare tutte le variabili in una sola tabella (**Confronta variabili**) o visualizzare una tabella di statistiche separata per ogni variabile (**Organizza l'output per variabili**).

Sopprimi le tabelle con molte categorie. 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 di riepilogo 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.

Considerazioni sui dati relative alla procedura Descrittive

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).

Ipotesi. 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

1. Dai menu, scegliere:
   Analizza > **Statistiche descrittive** > **Descrittive...**
2. 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

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 std.*. Una misura della dispersione intorno alla media. 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, l'età media è 45 e la deviazione standard è 10, il 95% dei casi sarà compreso tra 25 e 65 in un distribuzione normale.

- *Varianza*. Una misura della dispersione dei valori intorno alla media, pari alla somma dei quadrati delle deviazioni dalla media, divisa per un valore corrispondente al numero totale dei casi meno uno. La varianza è misurata in unità pari al quadrato di quelle della variabile stessa.
- *Intervallo*. La differenza tra il valore massimo ed il valore minimo di una variabile numerica, il massimo meno il minimo.
- *Minimo*. Il valore più basso di una variabile numerica.
- *Massimo*. Il valore più alto di una variabile numerica.
- *E. S. della media*. Una misura di quanto il valore della media può variare da campione a campione per campioni presi dalla stessa distribuzione. Può essere usata per comparare 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 raggruppino attorno a un punto centrale. 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 asimmetria positiva ha una 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 uno scostamento dalla 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`).

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

Capitolo 4. Esplora

La procedura Esplora produce statistiche di riepilogo 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 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 ritagliata al 5%, errore standard, varianza, deviazione standard, valore minimo e massimo, intervallo, intervallo interquartile, 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 diffusione contro intensità con test di Levene e trasformazioni.

Considerazioni sui dati relative alla procedura Esplora

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.

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

Per esplorare i dati

1. Dai menu, scegliere:
   Analizza > **Statistiche descrittive** > **Esplora...**
2. 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 delle frequenze.
- Fare clic su **Grafici** per ottenere istogrammi, grafici e test di probabilità normale e grafici di diffusione contro intensità con test di Levene.
- Fare clic su **Opzioni** per ottenere il trattamento dei valori mancanti.

Esplora: Statistica

Descrittive. Queste misure di tendenza centrale e di dispersione vengono visualizzate per impostazione predefinita. Le misure di tendenza centrale indicano l'ubicazione della distribuzione e includono la media, la mediana e la media ritagliata al 5%. Le misure di dispersione mostrano la dissimilarità dei valori e includono errore standard, varianza, deviazione standard, valore minimo e massimo, intervallo e intervallo interquartile. 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 dell'ubicazione. 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

Grafici a scatole. Queste alternative controllano la visualizzazione dei grafici a scatole quando sono presenti più variabili dipendenti. **Livelli dei fattori insieme** 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 delle probabilità normale e grafici delle probabilità normale detrendizzati. Viene visualizzato il test di Kolmogorov-Smirnov con un livello di significatività di Lilliefors per il test della normalità. Se sono specificati pesi non interi, la statistica di Shapiro-Wilk viene calcolata quando la dimensione del campione pesato è compresta tra 3 e 50. Per i pesi interi o in assenza di pesi, la statistica viene calcolata quando la dimensione del campione pesato è compresa tra 3 e 5.000.

Confronto tra diffusione e livello con test di Levene. Consente di controllare la trasformazione dati per i grafici di diffusione contro intensità. Per tutti i grafici di diffusione contro intensità vengono visualizzati la pendenza della linea 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 diffusione contro intensità. **Stima potenza** traccia i logaritmi naturali degli intervalli interquartili 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 diffusione 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. L'intervallo interquartile e la media dei dati trasformati verranno tracciati 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 dati

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

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

Esplora: Opzioni

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

- **Escludi casi a livello di elenco.** I casi con valori mancanti per qualsiasi variabile dipendente o fattore verranno esclusi da tutte le analisi. È l'impostazione predefinita.
- **Escludi casi a coppie.** 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.
- **Report valori.** I valori mancanti per le variabili fattore vengono trattati come categoria distinta. Tutto l'output viene prodotto per questa categoria supplementare. Le tabelle delle frequenze 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 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 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 diffusione vs. intensità (con il sottocomando `PLOT`).
- Specificare il numero di valori estremi da visualizzare (con il sottocomando `STATISTICS`).
- Specificare i parametri per gli stimatori M e gli stimatori robusti di ubicazione (con il sottocomando `MESTIMATORS`).

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

Capitolo 5. Tabelle di contingenza

La procedura Tabelle di contingenza consente di formare tabelle a due vie e a più dimensioni e fornisce una serie di test e misure di associazione per le tabelle a due vie. 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 tabelle di contingenza vengono calcolate solo per le tabelle a due vie. Se si specifica una riga, una colonna o un livello (variabile di controllo), verrà visualizzato un pannello contenente le statistiche associate e le misurazioni per ciascun valore del livello (o una combinazione di valori per due o più variabili di controllo). Ad esempio, se la variabile *sesso* è un livello per la tabella della variabile *coniugato* (sì, no) rispetto alla variabile *tipo di vita* (ottima, soddisfacente, non soddisfacente), i risultati per la tabella a due vie per le donne vengono elaborati separatamente da quelli per gli uomini e quindi stampati come pannelli 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 per lineare, test esatto di Fisher, chi-quadrato corretto di Yates, *r* di Pearson, rho di Spearman, coefficiente di contingenza, phi, *V* di Cramér, lambda simmetrico e asimmetrico, 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, odds ratio, test di McNemar, statistiche di Cochran e Mantel-Haenszel e statistiche delle proporzioni di colonna.

Considerazioni sui dati relativi alle tabelle di contingenza

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*.

Ipotesi. 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 un campione random di una distribuzione multinomiale.

Nota: le variabili ordinali possono essere codici numerici che rappresentano le categorie (ad esempio, 1 = *basso*, 2 = *medio*, 3 = *alto*) o valori 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

1. Dai menu, scegliere:
   Analizza > **Statistiche descrittive** > **Tabelle di contingenza...**
2. 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 a due vie.
- Fare clic su**Celle** per ottenere valori, percentuali e residui osservati e attesi.
- Fare clic su**Formato** per controllare l'ordine delle categorie.

Livello delle tabelle di contingenza

Se vengono selezionate una o più variabili di livello, verrà prodotta una tavola di contingenza distinta per ciascuna categoria di ciascuna variabile di livello (variabile di controllo). Ad esempio, se si dispone di una variabile di riga, una variabile di colonna e una variabile di livello con due categorie, si otterrà una tabella a due vie per ciascuna categoria della variabile di livello. Per creare un altro livello di variabili di controllo, fare clic su **Successivo**. Verranno create sottotabelle per ogni combinazione delle categorie di ciascuna variabile del primo livello con ciascuna variabile del secondo e così via. Se sono richieste statistiche e misure di associazione, verranno applicate solo alle sottotabelle a due vie.

Grafici a barre raggruppate di tabelle di contingenza

Visualizza grafici a barre raggruppate. Nei grafici a barre raggruppate 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 modelli diversi. Se in Colonne o Righe si specificano più variabili, verrà prodotto un grafico a barre raggruppate per ciascuna combinazione delle due variabili.

Tabelle di contingenza che visualizzano le variabili di livello nei livelli della tabella

Visualizza variabili di livello nei livelli della tabella. È possibile scegliere di visualizzare le variabili di livello (variabili di controllo) come livelli della tabella nella tabella delle tavole di contingenza. Ciò consente di creare delle viste che mostrano le statistiche generali per le variabili di riga e colonna e che consentono di eseguire il drill-down delle categorie delle variabili di livello.

Un esempio che utilizza il file di dati *demo.sav* (disponibile nella directory dei campioni della directory di installazione) viene visualizzato qui di seguito ed è stato ottenuto nel seguente modo:

1. Selezionare *Categoria di reddito in migliaia (inccat)* come variabile della riga, *Possiede PDA (ownpda)* come variabile della colonna e *Livello di istruzione (ed)* come variabile del livello.
2. Selezionare **Visualizza variabili di livello nei livelli della tabella**.
3. Selezionare **Colonna** nella finestra di dialogo secondaria Visualizzazione cella.
4. Eseguire la procedura Tabelle di contingenza, fare doppio clic sulla tabella delle tavole di contingenza e selezionare **Diploma di laurea** dall'elenco a discesa Livello di istruzione.

La vista selezionata della tabella delle tavole di contingenza mostra le statistiche per i rispondenti che hanno un diploma di laurea.

Statistiche delle tabelle di contingenza

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 che non risultata da righe o colonne mancanti in una tabella più grande, ha una cella con una frequenza prevista minore di 5. Il chi-quadrato corretto di Yates viene calcolato per tutte le altre tabelle 2×2. 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 le tabelle in cui sia le righe che le colonne contengono valori ordinati, l'opzione **Correlazioni** restituisce il coefficiente di correlazione di Spearman, rho (solo 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 o ebreo), è possibile selezionare **Coefficiente di contingenza**, **Phi** (coefficiente) **e V di Cramér**, **Lambda** (lambda simmetrici e asimmetrici e tau di Goodman e Kruskal) e **Coefficiente di incertezza**.

- *Coefficiente di contingenza.*. Una misura di associazione basata su 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 colonne in una tabella.
- *Phi e V di Cramer*. Phi è una misura di associazione che comporta la divisione della statistica di chi-quadrato per la dimensione del campione e l'estrazione della radice quadrata del risultato. V di Cramer è una misura di associazione basata sul chi-quadrato.
- *Lambda*. Una misura di associazione che riflette la riduzione proporzionale nell'errore quando i valori della variabile indipendente sono usati per prevedere i valori 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.*. Una misura di associazione che indica la riduzione proporzionale nell'errore quando i valori di una variabile vengono usati per prevedere i valori dell'altra variabile. 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. Il programma calcola sia la versione simmetrica che quella asimmetrica del coefficiente di incertezza.

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 va da -1 a 1. I valori prossimi a un 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. Per da 3 vie a n vie, vengono visualizzati i gamma condizionali.
- *D di Somers*. Una misura di associazione tra due variabili ordinali che va da -1 a 1. I valori prossimi a un valore assoluto di 1 indicano una forte relazione tra due variabili e i valori prossimi a 0 indicano una debole o inesistente relazione tra le variabili. È una estensione asimmetrica di gamma dalla quale differisce solo per l'inclusione del numero di coppie non correlate nella variabile indipendente. Viene calcolata anche una versione simmetrica di questa statistica.
- *Tau-b di Kendall*. Una misura non parametrica di correlazione per variabili ordinali o classificate che tiene conto delle correlazioni. Il segno del coefficiente indica la direzione della correlazione e il valore assoluto la sua forza. 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 non parametrica di associazione per variabili ordinali che ignora le correlazioni. Il segno del coefficiente indica la direzione della correlazione e il valore assoluto la sua forza. 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 compresa tra 0 e 1, dove 0 indica nessuna associazione tra le variabili riga e colonna e i valori prossimi a 1 indicano un alto grado 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 eta: il primo tratta la variabile riga come variabile di intervallo e l'altro tratta la variabile colonna come variabile di intervallo.

Kappa. Il kappa di Cohen misura l'accordo tra le valutazioni di due stimatori quando entrambi stanno dando un punteggio allo 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. Per le tabelle 2 x 2, una misura dell'intensità dell'associazione tra la presenza di un fattore e la ricorrenza di un evento. Un valore pari a 1 indica che il fattore non è associato all'evento. L'odds ratio può essere usato come una stima del rischio relativo quando la ricorrenza del fattore è rara.

McNemar. Un 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 progettazioni sperimentali del tipo 'prima e dopo', per rilevare cambiamenti di risposta. Per tabelle quadrate di dimensioni maggiori, viene notificato il test della simmetria di McNemar-Bowker.

Statistiche di Cochran e Mantel-Haenszel. Le statistiche di Cochran e Mantel-Haenszel possono essere usate per verificare l'indipendenza fra una variabile fattore dicotomica e una variabile di risposta dicotomica; la condizionalità è basata sui modelli di covariate definiti da una o più variabili di livello (controllo). Si noti che mentre le altre statistiche vengono calcolate livello per livello, le statistiche di Cochran e Mantel-Haenszel vengono calcolate una volta per tutti i livelli.

Visualizzazione celle delle tabelle di contingenza

Per facilitare l'individuazione di modelli di dati che danno origine a un test del chi-quadrato significativo, la procedura per le tabelle 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 che sono 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 superiore o uguale a 2, ma è consentito utilizzare il valore 0 per specificare che non viene nascosto alcun conteggio.

Confronta proporzioni di colonna. Questa opzione calcola le comparazioni a coppie delle proporzioni di colonna e indica quali coppie di colonne (per una determinata riga) sono differenti in modo significativo. Le differenze significative vengono indicate nella tabella delle tavole di contingenza con formattazione in stile APA utilizzando lettere come pedici e vengono calcolate al livello di significatività 0.05. *Nota*: se questa opzione viene specificata senza selezionare i conteggi osservati o le percentuali di colonna, i conteggi osservati vengono inclusi nella tabella delle tavole di contingenza con lettere come pedice in stile APA indicanti i risultati dei test delle proporzioni di colonna.

- **Adatta i valori P (metodo di Bonferroni).** Le comparazioni a coppie delle proporzioni di colonna utilizzano la correzione di Bonferroni, che adatta il livello di significatività osservato per il fatto che vengono eseguite comparazioni multiple.

Percentuali. Le percentuali possono essere aggiunte nelle righe o nelle colonne. Sono disponibili anche le percentuali del numero totale di casi rappresentati nella tabella (un solo livello). *Nota*: se nel gruppo Conteggi è 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 adattati.

- *Non standardizzati*. La differenza fra un valore osservato e il 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 sua deviazione standard. I residui standardizzati, noti anche come residui di Pearson, hanno una media pari a 0 e una deviazione standard pari a 1.
- *Standardizzati adattati*. Il residuo per una cella (valore osservato meno valore previsto) diviso per una stima del suo errore standard. Il residuo standardizzato risultante è espresso in unità di deviazione standard sopra o sotto la media.

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

- *Arrotonda conteggi delle celle*. I pesi del caso vengono usati così come sono ma prima del calcolo delle statistiche vengono arrotondati i pesi accumulati nelle celle.
- *Tronca conteggi delle celle*. I pesi del caso vengono usati così come sono, ma i pesi accumulati nelle celle vengono troncati prima di calcolare qualunque statistica.
- *Arrotonda pesi del caso*. I pesi del caso vengono arrotondati prima dell'uso.
- *Tronca pesi del caso*. I pesi del caso sono troncati prima dell'uso.
- *Nessun adattamento*. I pesi del caso vengono usati come sono e vengono usati anche i conteggi delle celle frazionari. Quando tuttavia è richiesto l'utilizzo dell'opzione Statistiche esatte (disponibile solo con l'opzione Test esatti), i pesi accumulati nelle celle vengono troncati o arrotondati prima del calcolo delle statistiche del test esatte.

Formato della tabella di tavole di contingenza

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

Capitolo 6. Riepiloga

La procedura Riassumi consente di calcolare la statistica del 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 di riepilogo. I valori dei dati di ciascuna categoria possono essere inseriti nell'elenco o eliminati, ma nei dataset 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 raggruppata, errore standard della media, minimo, massimo, intervallo, valore della variabile della prima categoria della variabile di raggruppamento, valore della variabile 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* totale, percentuale della somma in, percentuale del *numero di casi* in, media geometrica e media armonica.

Considerazioni sui dati relative alla procedura Riassumi

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.

Ipotesi. Alcune delle statistiche del 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 riepiloghi dei casi

1. Dai menu, scegliere:

   Analizza > **Report** > **Riepiloghi dei casi...**
2. 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

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 nei titoli e nelle didascalie digitando `\n` dove si desidera inserire un'interruzione della 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 identificativo nell'output.

Riassumi: Statistiche

È possibile scegliere una o più delle seguenti statistiche del sottogruppo per le variabili in ciascuna categoria di ogni variabile di raggruppamento: somma, numero di casi, media, mediana, mediana raggruppata, errore standard della media, minimo, massimo, intervallo, valore della variabile della prima categoria della variabile di raggruppamento, valore della variabile 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* totale, percentuale della somma in, percentuale del *numero di casi* in, media geometrica, media armonica. L'ordine in cui compaiono le statistiche nell'elenco Statistiche delle celle corrisponde all'ordine in cui verranno visualizzate nell'output. Per ciascuna variabile vengono visualizzate anche le statistiche di riepilogo in tutte le categorie.

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

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

Mediana raggruppata. La mediana calcolata per i dati codificati in gruppi. Ad esempio, con i dati di età, se ogni valore nella 30ina e codificato 35, ogni valore nella 40ina è codificato 45 e così via, la mediana raggruppata è la mediana calcolata dai dati codificati.

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

Curtosi. Una misura di quanto le osservazioni si raggruppino attorno a un punto centrale. 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 dei dati riscontrato nel file di dati.

Massimo. Il valore più alto di una variabile numerica.

Media. Una misura di tendenza centrale. La media aritmetica, ossia la somma divisa per il numero di casi.

Mediana. Il valore sopra il quale e sotto il quale ricade la metà dei casi, il 50° 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 estremamente bassi o alti.

Minimo. Il valore più basso di una variabile numerica.

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

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

Percentuale della somma totale.. Percentuale della somma totale in ciascuna categoria.

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

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 asimmetria positiva ha una 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 uno scostamento dalla simmetria.

Deviazione standard. Una misura della dispersione intorno alla media. 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, l'età media è 45 e la deviazione standard è 10, il 95% dei casi sarà compreso tra 25 e 65 in un distribuzione normale.

Errore standard della curtosi. Il rapporto della curtosi rispetto all'errore standard può essere usato come test di normalità (ovvero, è possibile rifiutare la normalità se il rapporto è inferiore a -2 o maggiore di +2). 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 (che diventano simili a quelle di una distribuzione uniforme a forma di scatola).

Errore standard della media. Una misura di quanto il valore della media può variare da campione a campione per campioni presi dalla stessa distribuzione. Può essere usata per comparare 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).

Errore standard dell'asimmetria. Il rapporto di asimmetria rispetto all'errore standard può essere usato come test di normalità (ovvero, è possibile rifiutare la normalità qualora il rapporto sia inferiore a -2 o maggiore di +2). 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 dei valori, su tutti i casi con valori non mancanti.

Varianza. Una misura della dispersione dei valori intorno alla media, pari alla somma dei quadrati delle deviazioni dalla media, divisa per un valore corrispondente al numero totale dei casi meno uno. La varianza è misurata in unità pari al quadrato di quelle della variabile stessa.

Capitolo 7. Medie

La procedura Medie consente di calcolare le medie dei sottogruppi e la statistica univariata correlata per le variabili dipendenti all'interno delle categorie di una o più variabili indipendenti. È inoltre possibile ottenere l'analisi della varianza a una via, eta e test di linearità.

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

Statistiche. Somma, numero di casi, media, mediana, mediana raggruppata, errore standard della media, minimo, massimo, intervallo, valore della variabile della prima categoria della variabile di raggruppamento, valore della variabile 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* totale, percentuale della somma in, percentuale del *numero di casi* in, media geometrica e media armonica. Le opzioni includono analisi della varianza, eta, eta al quadrato, e test di linearità R e R^2.

Considerazioni sui dati relativi alle medie

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

Ipotesi. Alcune delle statistiche del 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 a una via.

Per ottenere le medie dei sottogruppi

1. Dai menu, scegliere:

   Analizza > **Confronta medie** > **Medie...**
2. Selezionare una o più variabili dipendenti.
3. 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 livelli di variabili indipendenti. Ogni livello suddivide ulteriormente il campione. Se è presente una sola variabile indipendente nello Livello 1 e una sola nello Livello 2, i risultati verranno visualizzati in una tabella incrociata e non in tabelle distinte per ciascuna variabile indipendente.
4. Oppure fare clic su **Opzioni** per ottenere statistiche facoltative, analisi della tabella di varianza, età, età quadrato, R e R^2.

Medie: Opzioni

È possibile scegliere una o più delle seguenti statistiche del sottogruppo per le variabili in ciascuna categoria di ogni variabile di raggruppamento: somma, numero di casi, media, mediana, mediana raggruppata, errore standard della media, minimo, massimo, intervallo, valore della variabile della prima categoria della variabile di raggruppamento, valore della variabile 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* totale, percentuale della somma in, percentuale del *numero di casi* in, media geometrica e media armonica. È possibile modificare l'ordine in cui compare la statistica del sottogruppo. L'ordine in cui compaiono le statistiche nell'elenco Statistiche delle celle corrisponde all'ordine in cui verranno visualizzate nell'output. Per ciascuna variabile vengono visualizzate anche le statistiche di riepilogo in tutte le categorie.

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

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

Mediana raggruppata. La mediana calcolata per i dati codificati in gruppi. Ad esempio, con i dati di età, se ogni valore nella 30ina e codificato 35, ogni valore nella 40ina è codificato 45 e così via, la mediana raggruppata è la mediana calcolata dai dati codificati.

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

Curtosi. Una misura di quanto le osservazioni si raggruppino attorno a un punto centrale. 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 dei dati riscontrato nel file di dati.

Massimo. Il valore più alto di una variabile numerica.

Media. Una misura di tendenza centrale. La media aritmetica, ossia la somma divisa per il numero di casi.

Mediana. Il valore sopra il quale e sotto il quale ricade la metà dei casi, il 50° 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 estremamente bassi o alti.

Minimo. Il valore più basso di una variabile numerica.

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

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

Percentuale della somma totale. Percentuale della somma totale in ciascuna categoria.

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

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 asimmetria positiva ha una 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 uno scostamento dalla simmetria.

Deviazione standard. Una misura della dispersione intorno alla media. 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, l'età media è 45 e la deviazione standard è 10, il 95% dei casi sarà compreso tra 25 e 65 in un distribuzione normale.

Errore standard della curtosi. Il rapporto della curtosi rispetto all'errore standard può essere usato come test di normalità (ovvero, è possibile rifiutare la normalità se il rapporto è inferiore a -2 o maggiore di +2). 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 (che diventano simili a quelle di una distribuzione uniforme a forma di scatola).

Errore standard della media. Una misura di quanto il valore della media può variare da campione a campione per campioni presi dalla stessa distribuzione. Può essere usata per comparare 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).

Errore standard dell'asimmetria. Il rapporto di asimmetria rispetto all'errore standard può essere usato come test di normalità (ovvero, è possibile rifiutare la normalità qualora il rapporto sia inferiore a -2 o maggiore di +2). 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 dei valori, su tutti i casi con valori non mancanti.

Varianza. Una misura della dispersione dei valori intorno alla media, pari alla somma dei quadrati delle deviazioni dalla media, divisa per un valore corrispondente al numero totale dei casi meno uno. La varianza è misurata in unità pari al quadrato di quelle della variabile stessa.

Statistiche per il primo livello

Tabella ANOVA ed Eta. Visualizza una tabella di analisi della varianza a una via e calcola le misure di associazione eta ed eta al quadrato per ogni variabile indipendente nel primo livello.

Test di linearità. Calcola la somma dei quadrati, i gradi di libertà e la media quadratica associati 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 di riepilogo continue all'interno delle categorie di una o più variabili di raggruppamento categoriali. Nella tabella viene creato un livello 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.

Considerazioni sui dati dei cubi OLAP

Dati. Le variabili di riepilogo 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.

Ipotesi. Alcune delle statistiche del 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

1. Dai menu, scegliere:

   Analizza > **Report** > **Cubi OLAP...**
2. Selezionare una o più variabili di riepilogo continue.
3. Selezionare una o più variabili categoriali di raggruppamento

Oppure:

- Selezionare statistiche di riepilogo diverse (fare clic su **Statistiche**). Prima di selezionare le statistiche di riepilogo 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 che sono 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 superiore o uguale a 2.

Consultare per informazioni su come vengono rese le dimensioni di più livelli nelle tabelle semplici.

Cubi OLAP: Statistiche

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

È possibile modificare l'ordine in cui compare la statistica del sottogruppo. L'ordine in cui compaiono le statistiche nell'elenco Statistiche delle celle corrisponde all'ordine in cui verranno visualizzate nell'output. Per ciascuna variabile vengono visualizzate anche le statistiche di riepilogo in tutte le categorie.

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

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

Mediana raggruppata. La mediana calcolata per i dati codificati in gruppi. Ad esempio, con i dati di età, se ogni valore nella 30ina e codificato 35, ogni valore nella 40ina è codificato 45 e così via, la mediana raggruppata è la mediana calcolata dai dati codificati.

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

Curtosi. Una misura di quanto le osservazioni si raggruppino attorno a un punto centrale. 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 dei dati riscontrato nel file di dati.

Massimo. Il valore più alto di una variabile numerica.

Media. Una misura di tendenza centrale. La media aritmetica, ossia la somma divisa per il numero di casi.

Mediana. Il valore sopra il quale e sotto il quale ricade la metà dei casi, il 50° 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 estremamente bassi o alti.

Minimo. Il valore più basso di una variabile numerica.

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

Percentuale del numero di casi in. Percentuale del numero di casi per la variabile di raggruppamento specificata entro le categorie di altre variabili di raggruppamento. Se si ha una sola variabile di raggruppamento, questo valore è identico alla percentuale del numero totale di casi.

Percentuale della somma in. Percentuale della somma per la variabile di raggruppamento specificata entro le categorie di altre variabili di raggruppamento. Se si ha una sola variabile di raggruppamento, questo valore è identico alla percentuale della somma totale.

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

Percentuale della somma totale.. Percentuale della somma totale in ciascuna categoria.

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

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 asimmetria positiva ha una 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 uno scostamento dalla simmetria.

Deviazione standard. Una misura della dispersione intorno alla media. 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, l'età media è 45 e la deviazione standard è 10, il 95% dei casi sarà compreso tra 25 e 65 in un distribuzione normale.

Errore standard della curtosi. Il rapporto della curtosi rispetto all'errore standard può essere usato come test di normalità (ovvero, è possibile rifiutare la normalità se il rapporto è inferiore a -2 o maggiore di +2). 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 (che diventano simili a quelle di una distribuzione uniforme a forma di scatola).

Errore standard della media. Una misura di quanto il valore della media può variare da campione a campione per campioni presi dalla stessa distribuzione. Può essere usata per comparare 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).

Errore standard dell'asimmetria. Il rapporto di asimmetria rispetto all'errore standard può essere usato come test di normalità (ovvero, è possibile rifiutare la normalità qualora il rapporto sia inferiore a -2 o maggiore di +2). 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 dei valori, su tutti i casi con valori non mancanti.

Varianza. Una misura della dispersione dei valori intorno alla media, pari alla somma dei quadrati delle deviazioni dalla media, divisa per un valore corrispondente al numero totale dei casi meno uno. La varianza è misurata in unità pari al quadrato di quelle della variabile stessa.

Cubi OLAP: Differenze

Questa finestra di dialogo consente di calcolare le differenze aritmetiche e percentuali tra variabili di riepilogo 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 di riepilogo della seconda variabile di ogni coppia (la variabile sottratta) vengono sottratti dai valori delle statistiche di riepilogo della prima variabile della coppia. Per le differenze percentuali, il valore della variabile di riepilogo della variabile sottratta viene utilizzato al denominatore. È necessario selezionare almeno due variabili di riepilogo 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 di riepilogo della seconda categoria di ogni coppia (la categoria sottratta) vengono sottratti dai valori delle statistiche di riepilogo della prima categoria della coppia. Per le differenze percentuali, il valore delle statistiche di riepilogo 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

È possibile modificare il titolo dell'output o aggiungere una didascalia che verrà visualizzata sotto la tabella dell'output. È anche possibile controllare gli a capo automatici nei titoli e nelle didascalie digitando `\n` dove si desidera inserire un'interruzione della riga nel testo.

Capitolo 9. Test T

Test T

Sono disponibili tre tipi di test *t*:

Test T: campioni indipendenti (test T: 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: campioni accoppiati (test T: dipendente). Consente di confrontare le medie di due variabili per un singolo gruppo. Questo test viene utilizzato anche per progettazioni relative 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: campione unico. 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.

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 di trattamento. La pressione di ogni paziente viene misurata una volta e ciascun paziente appartiene a un solo gruppo.

Statistiche. Per ogni variabile: dimensione del campione, media, deviazione standard ed errore standard della media. Per la differenza nelle medie: media, errore standard e intervallo di confidenza (è possibile specificare il livello di confidenza). Test: test di Levene per l'eguaglianza delle varianze ed entrambi i test di varianze raggruppate e varianze separate *t* per l'eguaglianza delle medie.

Considerazioni sui dati di test T per campioni indipendenti

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).

Ipotesi. Per il test *t* di uguaglianza della varianza, le osservazioni dovrebbero essere rappresentate da campioni random e indipendenti derivati da distribuzioni normali con la stessa varianza di popolazione. Per il test *t* di inuguaglianza della varianza, le osservazioni dovrebbero essere campioni random e indipendenti 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

1. Dai menu, scegliere:

   Analizza > **Confronta medie** > **Test T: campioni indipendenti...**
2. Selezionare una o più variabili quantitative oggetto del test. Per ciascuna variabile viene calcolato un test *t* distinto.
3. Selezionare una variabile di raggruppamento singola e fare clic su **Definisci gruppi** per specificare due codici per i gruppi che si desidera confrontare.
4. 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

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.** Immettere 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 al punto di divisione da un gruppo e i casi con valori maggiori o uguali al punto di divisione dall'altro gruppo.

Per le variabili di raggruppamento di tipo stringa, immettere una stringa per Gruppo 1 e un altro valore per Gruppo 2, ad esempio, *sì* e *no*. I casi con altre stringhe verranno esclusi dall'analisi.

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 mancati. 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):

- **Escludi 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.
- **Escludi casi a livello di elenco.** 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.

Test T per campioni accoppiati

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 progettazioni relative 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 del campione, deviazione standard ed errore standard della media. Per ogni coppia di variabili: correlazione, differenza media nelle medie, test *t* e intervallo di confidenza per la differenza della media (è possibile specificare il livello di confidenza). Deviazione standard ed errore standard della differenza media.

Considerazioni sui dati di test T per campioni appaiati

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.

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

Per ottenere un test T per campioni appaiati

1. Dai menu, scegliere:

   Analizza > **Confronta medie** > **Test T: campioni accoppiati...**
2. Selezionare una o più coppie di variabili
3. 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

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 mancati. 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):

- **Escludi 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.
- **Escludi casi a livello di elenco.** 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.

Funzioni 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`).

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

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 voler testare se il punteggio IQ medio di un gruppo di stundenti è diverso da 100. Oppure un produttore di cereali può prendere un campione di scatole dalla linea di produzione e controllare se il peso medio dei campioni è diverso da 1.3 libre al livello di confidenza del 95%.

Statistiche. Per ogni variabile di test: media, deviazione standard ed errore standard della media. La differenza media fra ciascun valore dei dati 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).

Considerazioni sui dati di test T per un campione

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.

Ipotesi. 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

1. Dai menu, scegliere:

   Analizza > **Confronta medie** > **Test T: campione unico...**
2. Selezionare una o più variabili da confrontare con lo stesso valore ipotizzato.
3. Immettere un valore oggetto del test numerico rispetto al quale viene confrontata ciascuna media del campione.
4. 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

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 mancati. 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):

- **Escludi 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.
- **Escludi casi a livello di elenco.** 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.

Funzioni 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`).

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

Funzioni 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`).

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

Capitolo 10. ANOVA a una via

La procedura ANOVA a una via produce un'analisi della varianza a una via 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. Sono disponibili due tipi di test per il confronto delle 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 le tendenze presenti tra le categorie.

Esempio. Le ciambelle assorbono quantità variabili di grassi a seconda della modalità di cottura. È stato impostato un esperimento che coinvolge 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 ogni gruppo: numero di casi, media, deviazione standard, errore standard della media, valore minimo e massimo e intervallo di confidenza del 95% per la media. Test di Levene per l'omogeneità di varianza, tabella di analisi della varianza e test robusti dell'uguaglianza delle medie per ogni variabile dipendente, contrasti a priori specificati dall'utente, test dell'intervallo post hoc e comparazioni multiple: Bonferroni, Sidak, differenza significativa di Tukey, GT2 di Hochberg, Gabriel, Dunnett, test *F* di Ryan-Einot-Gabriel-Welsch (*F* di R-E-G-W), test dell'intervallo di Ryan-Einot-Gabriel-Welsch (*Q* di R-E-G-W), T2 di Tamhane, T3 di Dunnett, Games-Howell, *C* di Dunnett, test di intervallo multiplo di Duncan, Student-Newman-Keuls (S-N-K), *b* di , Waller-Duncan, Scheffé e differenza meno significativa.

Considerazioni sui dati relativi all'ANOVA a una via

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

Ipotesi. Ciascun gruppo è un campione random 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 a una via

1. Dai menu, scegliere:
   Analizza > **Confronta medie** > **ANOVA a una via...**
2. Selezionare una o più variabili dipendenti.
3. Selezionare una singola variabile fattore indipendente.

ANOVA a una via: Contrasti

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

Polinomiale. Consente di suddividere le somme dei quadrati tra gruppi in componenti di tendenza. È possibile verificare una tendenza della variabile dipendente in tutti i livelli ordinati della variabile fattore. Ad esempio, è possibile verificare una tendenza 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 esistono sei categorie della variabile fattore, i coefficienti –1, 0, 0, 0, 0.5 e 0.5 contrastano il primo gruppo con i gruppi quinto e sesto. 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 avviso.

ANOVA a una via: Test Post Hoc

Dopo aver determinato l'esistenza di differenze tra le medie, i test post hoc di intervallo e le comparazioni a coppie multiple consentono di determinare quale media differisce dalle altre. I test di intervallo multipli consentono di identificare sottoinsiemi omogenei di medie che non differiscono le une dalle altre. Grazie alle comparazioni a coppie multiple 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.

Presumi varianze uguali

Test Differenza significativa di Tukey, GT2 di Hochberg, Gabriel e Scheffé sono test di comparazione multipla e test di intervallo. Sono disponibili altri test di intervallo, 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 di intervallo di Ryan-Einot-Gabriel-Welsch) e Waller-Duncan. I test di comparazione multipla disponibili sono Bonferroni, il test Differenza significativa di Tukey, Sidak, Gabriel, Hochberg, Dunnett, Scheffé e LSD (least significant difference).

- *LSD*. Usa i test t per eseguire tutte le comparazioni a coppie tra medie di gruppo. Non viene apportata alcuna correzione al tasso di errore per comparazioni multiple.
- *Bonferroni*. Usa i test t per eseguire comparazioni 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. Il livello di significatività osservato viene pertanto è adattato tenendo conto che si stanno effettuando comparazioni multiple.
- *Sidak*. Test di comparazione multipla a coppie basato su una statistica t. Sidak corregge il livello di significatività per comparazioni multiple e fornisce dei limiti più stretti del Bonferroni.
- *Scheffé*. Effettua comparazioni congiunte a coppie simultanee per tutte le possibili combinazioni a coppie di medie. Usa la distribuzione di campionamento F. Può essere usato per esaminare tutte le possibili combinazioni lineari di medie di gruppo, non solo le comparazioni a coppie.
- *R-E-G-W F*. La procedura (a multipli) decrescenti di Ryan-Einot-Gabriel-Welsch, basata su un test F.
- *R-E-G-W Q*. La procedura (a multipli) decrescenti di Ryan-Einot-Gabriel-Welsch, basata sull'intervallo studentizzato.
- *S-N-K*. Effettua tutte le comparazioni a coppie fra medie usando la distribuzione di intervallo studentizzata. Per dimensioni del campione uguali, confronta anche coppie di medie entro sottoinsiemi omogenei, usando una procedura a fasi. 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 tutte le comparazioni a coppie tra gruppi. Imposta il tasso di errore sperimentale al valore del tasso di errore per l'insieme di tutte le comparazioni a coppie.
- *b di Tukey*. Usa la distribuzione di intervallo studentizzato per effettuare comparazioni a coppie tra gruppi. Il valore critico è la media fra il corrispondente valore per il test HSD (honestly significant difference) di Tukey e quello di Student-Newman-Keuls.
- *Duncan*. Esegue le comparazioni a coppie usando un ordine di comparazioni a fasi identico a quello usato dal test di Student-Newman-Keuls, ma imposta un livello di protezione per il tasso di errore per la raccolta dei test piuttosto che un tasso di errore per i singoli test. Usa la statistica dell'intervallo studentizzato.
- *GT2 di Hochberg*. Test di comparazione multipla e di intervallo basato sul modulo massimo studentizzato. È simile al test HSD (honestly significant difference) di Tukey.
- *Gabriel*. Test di comparazione a coppie basato sul modulo massimo studentizzato, generalmente più potente del GT2 di Hochberg 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 di comparazione multipla basato su una statistica t; usa un approccio bayesiano.
- *Dunnett*. Test t di comparazione multipla a coppie che compara una serie di trattamenti a una singola 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 presumere varianze uguali

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

- *T2 di Tamhane*. Test di comparazioni a coppie conservativi basati su un test t. Questo test è appropriato quando le varianze non sono uguali.
- *T3 di Dunnett*. Test di comparazione a coppie basato sul modulo massimo studentizzato. Questo test è appropriato quando le varianze non sono uguali.
- *Games-Howell*. Test di comparazione a coppie che a volte non è molto conservativo. Questo test è appropriato quando le varianze non sono uguali.
- *C di Dunnett*. Test di comparazione a coppie basato sull'intervallo studentizzato. Questo test è appropriato quando le varianze non sono uguali.

Nota: può essere più facile interpretare l'output dei test post hoc se si deseleziona **Nascondi righe e colonne vuote** nella finestra di dialogo Proprietà tabella (in una tabella pivot attivata scegliere **Proprietà tabella** dal menu Formato).

ANOVA a una via: 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 random.** 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 random.
- **Test di omogeneità della 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.** Calcola la statistica di Brown-Forsythe per eseguire il test di eguaglianza delle medie di gruppi. Questa statistica è preferibile alla statistica *F* nel caso in cui non sia valida l'ipotesi di uguaglianza della varianza.
- **Welch.** Calcola la statistica di Welch per eseguire il test di eguaglianza delle medie di gruppi. Questa statistica è preferibile alla statistica *F* nel caso in cui non sia valida l'ipotesi di uguaglianza della varianza.

Grafico delle medie. Visualizza un grafico che traccia le medie del sottogruppo (le medie per ogni gruppo definito dai valori della variabile fattore).

Valori mancanti. Controlla come vengono considerati i valori mancanti.

- **Escludi casi analisi per analisi.** Un caso con un valore mancante per la variabile dipendente o fattore per una determinata analisi non viene utilizzato nell'analisi in questione. Non verranno utilizzati nemmeno i casi che non rientrano nell'intervallo specificato per la variabile fattore.
- **Escludi casi a livello di elenco.** I casi con valori mancanti per la variabile fattore o per una qualsiasi variabile dipendente inclusa nell'elenco di variabili dipendenti nella finestra di dialogo principale vengono esclusi da tutte le analisi. Se non sono state specificate più variabili dipendenti, l'opzione non produrrà alcun effetto.

Funzioni aggiuntive del comando ONEWAY

Il linguaggio della sintassi dei comandi consente inoltre di:

- Ottenere statistiche con effetti fissi e random. 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 random (con il sottocomando `STATISTICS=EFFECTS`).
- Specificare i livelli alfa per la differenza meno significativa, i test di comparazione multipla 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 l'analisi della varianza a una via (con il sottocomando `MATRIX`).

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

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 null 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 random. 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. Una progettazione è bilanciata 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 di leva possono essere salvati come variabili nel file di dati per la verifica di ipotesi.

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

Esempio. Vengono raccolti i dati per i singoli partecipanti della maratona di Chicago per diversi anni. 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. Le somme dei quadrati di tipo I, tipo II, tipo III e tipo IV possono essere utilizzate per valutare ipotesi differenti. Il metodo predefinito è il Tipo III.

Statistiche. Test post hoc di intervallo e comparazioni multiple: differenze meno significative, Bonferroni, Sidak, Scheffé, *F* multiplo di Ryan-Einot-Gabriel-Welsch, intervallo multiplo di Ryan-Einot-Gabriel-Welsch, Student-Newman-Keuls, differenza significativa di Tukey, *b* di Tukey, Duncan, GT2 di Hochberg, Gabriel, test *t* di Waller-Duncan, Dunnett (a 1 via e a 2 vie), T2 di Tamhane, T3 di Dunnett, Games-Howell e *C* di Dunnett. Statistiche descrittive: medie osservate, deviazioni standard e conteggi per tutte le variabili dipendenti in tutte le celle. Test di Levene per l'omogeneità della varianza.

Grafici. Di confronto tra diffusione e livello, residuo e di profilo (interazione).

Considerazioni sui dati di GLM univariato

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.

Ipotesi. I dati sono costituiti da un campione random 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 diffusione vs. intensità. È inoltre possibile esaminare residui e grafici dei residui.

Per ottenere tabelle di GLM univariato

1. Dai menu, scegliere:

   Analizza > **Modello lineare generale** > **Univariata...**
2. Selezionare una variabile dipendente.
3. Selezionare le variabili per l'opzione Fattori fissi, Fattori random o Covariate, a seconda dei dati in uso.
4. È inoltre possibile utilizzare l'opzione Peso WLS per specificare una variabile peso per l'analisi dei minimi quadrati pesati. Casi con valore 0, negativo o mancante per la variabile peso saranno esclusi dall'analisi. Non è possibile utilizzare come variabile peso una variabile già inclusa nel modello.

GLM – Univariato: Modello

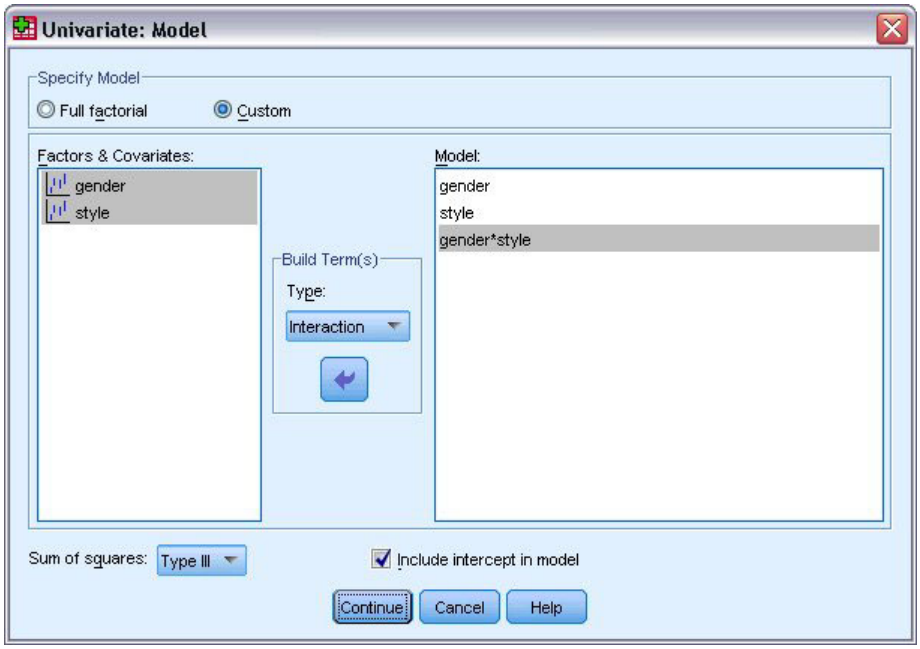


Figura 1. Finestra di dialogo Univariata: Modello

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 dei 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 della somma dei quadrati in genere utilizzato con modelli bilanciati o non bilanciati privi di celle mancanti è il Tipo III.

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

Crea 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 noto anche come decomposizione gerarchica del metodo Somma dei quadrati. Ciascun termine viene adattato 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 adattato 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 della 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.
- Una progettazione nidificata in modo puro. Questo tipo di nidificazione può essere specificato tramite la sintassi.

Tipo III. Tipo predefinito. Questo metodo calcola le somma dei quadrati di un effetto nella progettazione come la somma dei quadrati adattata per qualsiasi altro effetto che non lo contiene 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 una progettazione fattoriale priva di celle mancanti, questo metodo equivale alla tecnica dei quadrati delle medie pesate di Yates. Il metodo della 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* della progettazione, 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 della 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

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.

GLM univariato. Il test sull'ipotesi è basato sull'ipotesi null **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 adattate 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 dei fattori 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 delle tendenze polinomiali.

GLM – Univariato: Grafici di profilo

I grafici di profilo, o grafici di interazione, risultano utili per il confronto delle medie marginali di un modello. Un grafico di profilo è un grafico a linee in cui ciascun punto indica la media marginale stimata di una variabile dipendente (adattata 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 random 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 entro soggetti che fattori tra soggetti. GLM Multivariato e GLM Misure ripetute sono disponibili solo se è stata installata l'opzione Advanced Statistics.

Il grafico di 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.

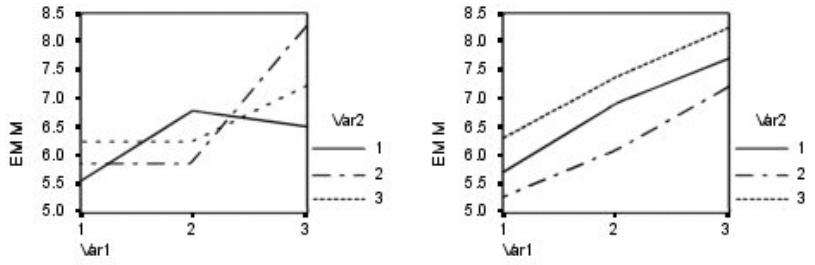


Figura 2. Grafico con linee non parallele (a sinistra) e grafico con linee parallele (a destra)

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: 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 adattate per le eventuali covariate.

- **Confronta effetti principali.** Consente di eseguire comparazioni 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.
- **Adattamento intervallo di confidenza.** Selezionare la differenza meno significativa (LSD), l'adattamento di Bonferroni o di Sidak agli intervalli di confidenza e la significatività. Questo comando è disponibile solo se è stato selezionato **Confronta effetti principali**.

Visualizza. Selezionare **Statistica descrittiva** 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 al quadrato parziale per ciascun effetto e per ciascuna stima del parametro. La statistica eta al 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 del parametro** per ottenere stime del parametro, 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 tra soggetti, solo per i fattori tra soggetti. Le opzioni Grafici di diffusione 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 se la relazione tra la variabile dipendente e le variabili indipendenti può essere descritta in modo adeguato dal modello. La funzione **Forma funzionale generalizzata** consente di creare test sull'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 adattare 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 della progettazione (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`).
- Creare una matrice **L**, una matrice **M** o una matrice **K** personalizzata (utilizzando i sottocomandi `LMATRIX`, `MMATRIX` e `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 comparazioni post-hoc (tramite il sottocomando `POSTHOC`).
- Calcolare medie marginali stimate per qualsiasi fattore o interazione tra fattori nell'elenco dei fattori (tramite il sottocomando `EMMEANS`).
- Assegnare un nome alle variabili temporanee (tramite il sottocomando `SAVE`).
- Costruire un file di dati matrice di correlazione (tramite il sottocomando `OUTFILE`).
- Costruire un file di dati matrice contenente statistiche derivate dai dati della tabella ANOVA tra soggetti (tramite il sottocomando `OUTFILE`).
- Salvare la matrice di progettazione in un nuovo file di dati (tramite il sottocomando `OUTFILE`).

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

GLM: Comparazioni post hoc

Test di comparazioni multiple post hoc. Dopo aver determinato l'esistenza di differenze tra le medie, i test post hoc di intervallo e le comparazioni a coppie multiple consentono di determinare quale media differisce dalle altre. Le comparazioni vengono eseguite su valori a cui non è stato apportato alcun adattamento. 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 di comparazione multipla 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 di comparazione multipla usati più di frequente sono il test di Bonferroni e il test HSD di Tukey. Il **test di Bonferroni**, basato sulla statistica *t* di Student, consente di adattare il livello di significatività osservato in base al fatto che vengono eseguite comparazioni multiple. Il **test t di Sidak** adatta inoltre il livello di significatività ed è più restrittivo del test di Bonferroni. Il **test HSD di Tukey** utilizza la statistica di intervallo studentizzato per effettuare tutte le comparazioni a coppie tra gruppi e imposta il tasso di errore sperimentale sul valore del tasso di errore per l'insieme di tutte le comparazioni a coppie. Quando si eseguono test su un elevato numero di coppie di medie, il test HSD 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 HSD di Tukey, ma utilizza il modulo massimo studentizzato. Il test di Tukey risulta in genere più efficace. Anche il **test delle comparazioni 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 comparazioni multiple 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 dei fattori (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 di intervallo 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 di 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, utilizzare **T2 di Tamhane** (test di comparazioni a coppie conservativo basato su un test *t*), **T3 di Dunnett** (test di comparazione a coppie basato sui moduli massimi studentizzati), **il test di comparazione a coppie di Games-Howell** (a volte liberale) o **C di Dunnett** (test di comparazione 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 di intervallo multiplo di Duncan, il test di Student-Newman-Keuls (**S-N-K**) e il test **b di Tukey** sono test di 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 di intervallo che usa la media armonica della dimensione del campione nel caso di dimensioni del campione non uguali.

Il livello di significatività del test di **Scheffé** è progettato per consentire il test di tutte le possibili combinazioni lineari delle medie di gruppo, non solo delle comparazioni a coppie disponibili in questa funzione. Ne risulta che il test di Scheffé è spesso più conservativo di altri test, ciò significa che è richiesta una maggiore differenza tra le medie per la significatività.

Il test di comparazione multipla a coppie 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 adattamento del livello di significatività osservata per comparazioni multiple.

Test visualizzati. Le comparazioni a coppie sono disponibili per i test LSD, Sidak, Bonferroni, Games-Howell, Tamhane (T2) e (T3), *C* di Dunnett e T3 di Dunnett. 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 di intervallo. Il test Differenza significativa di Tukey, GT2 di Hochberg, il test di Gabriel e il test di Scheffé sono sia test di comparazione multipla che test di intervalli.

GLM – Univariato: 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 adattate per le eventuali covariate.

- **Confronta effetti principali.** Consente di eseguire comparazioni 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.
- **Adattamento intervallo di confidenza.** Selezionare la differenza meno significativa (LSD), l'adattamento di Bonferroni o di Sidak agli intervalli di confidenza e la significatività. Questo comando è disponibile solo se è stato selezionato **Confronta effetti principali**.

Visualizza. Selezionare **Statistica descrittiva** 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 al quadrato parziale per ciascun effetto e per ciascuna stima del parametro. La statistica eta al 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 del parametro** per ottenere stime del parametro, 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 tra soggetti, solo per i fattori tra soggetti. Le opzioni Grafici di diffusione 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 se la relazione tra la variabile dipendente e le variabili indipendenti può essere descritta in modo adeguato dal modello. La funzione **Forma funzionale generalizzata** consente di creare test sull'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 adattare 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 della progettazione (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).
- Creare una matrice **L**, una matrice **M** o una matrice **K** personalizzata (utilizzando i sottocomandi LMATRIX, MMATRIX e 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 comparazioni post-hoc (tramite il sottocomando POSTHOC).
- Calcolare medie marginali stimate per qualsiasi fattore o interazione tra fattori nell'elenco dei fattori (tramite il sottocomando EMMEANS).
- Assegnare un nome alle variabili temporanee (tramite il sottocomando SAVE).
- Costruire un file di dati matrice di correlazione (tramite il sottocomando OUTFILE).
- Costruire un file di dati matrice contenente statistiche derivate dai dati della tabella ANOVA tra soggetti (tramite il sottocomando OUTFILE).
- Salvare la matrice di progettazione in un nuovo file di dati (tramite il sottocomando OUTFILE).

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

GLM – Univariato: 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 previsti. Valori attesi dal modello per ciascun caso.

- *Non standardizzati*. Il valore previsto dal modello per la variabile dipendente.
- *Pesato*. Valori previsti non standardizzati pesati. Disponibile solo se era stata precedentemente 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.

Diagnostiche. 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. Un valore alto del D di Cook indica che l'esclusione di un caso dal calcolo delle statistiche di regressione modifica sostanzialmente i coefficienti.
- *Valori di leva*. Valori di leva non centrati. L'influenza relativa di ogni osservazione sull'adattamento del modello.

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 previsto dal modello.
- *Pesato*. Residui non standardizzati pesati. Disponibile solo se era stata precedentemente selezionata una variabile WLS.
- *Standardizzati*. Il residuo diviso per una stima della sua deviazione standard. I residui standardizzati, noti anche come residui di Pearson, hanno una media pari a 0 e una deviazione standard pari a 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 ciascun caso nelle variabili indipendenti e le medie delle variabili indipendenti.

- *Eliminato*. Il residuo per un caso quando tale caso viene escluso dal calcolo dei coefficienti di regressione. È la differenza tra il valore della variabile dipendente e il valore previsto adattato.

Statistiche dei coefficienti. Scrive una matrice della varianza-covarianza delle stime del parametro nel modello in un nuovo dataset della sessione attiva o in un file di dati IBM SPSS Statistics esterno. Per ciascuna variabile dipendente è inoltre disponibile una riga di stime del parametro, una riga di valori di significatività per le statistiche *t* corrispondenti alle stime del parametro 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

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 adattate per le eventuali covariate.

- **Confronta effetti principali.** Consente di eseguire comparazioni 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.
- **Adattamento intervallo di confidenza.** Selezionare la differenza meno significativa (LSD), l'adattamento di Bonferroni o di Sidak agli intervalli di confidenza e la significatività. Questo comando è disponibile solo se è stato selezionato **Confronta effetti principali**.

Visualizza. Selezionare **Statistica descrittiva** 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 al quadrato parziale per ciascun effetto e per ciascuna stima del parametro. La statistica eta al 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 del parametro** per ottenere stime del parametro, 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 tra soggetti, solo per i fattori tra soggetti. Le opzioni Grafici di diffusione 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 se la relazione tra la variabile dipendente e le variabili indipendenti può essere descritta in modo adeguato dal modello. La funzione **Forma funzionale generalizzata** consente di creare test sull'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 adattare 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 della progettazione (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).
- Creare una matrice **L**, una matrice **M** o una matrice **K** personalizzata (utilizzando i sottocomandi LMATRIX, MMATRIX e 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 comparazioni post-hoc (tramite il sottocomando POSTHOC).
- Calcolare medie marginali stimate per qualsiasi fattore o interazione tra fattori nell'elenco dei fattori (tramite il sottocomando EMMEANS).
- Assegnare un nome alle variabili temporanee (tramite il sottocomando SAVE).
- Costruire un file di dati matrice di correlazione (tramite il sottocomando OUTFILE).
- Costruire un file di dati matrice contenente statistiche derivate dai dati della tabella ANOVA tra soggetti (tramite il sottocomando OUTFILE).
- Salvare la matrice di progettazione in un nuovo file di dati (tramite il sottocomando OUTFILE).

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

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. Analizzando i dati della stagione NBA 1994–1995 si deduce che il coefficiente di correlazione di Pearson (0.581) è significativo a 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 correlate in modo negativo (–0.401) e la correlazione è significativa a livello 0.05.

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

Considerazioni sui dati relativi alle correlazioni bivariate

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.

Ipotesi. 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...**

1. Selezionare una o più variabili numeriche.

Sono inoltre disponibili le seguenti opzioni:

- **Coefficienti di correlazione.** Per le 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. Il valore dei coefficienti di correlazione è compreso tra –1 (una relazione negativa perfetta) e +1 (una 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 due code o a una coda. 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

Statistiche. Per le correlazioni Pearson 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. I valori mancanti vengono gestiti variabile per variabile indipendentemente dall'impostazione corrispondente.
- **Deviazioni e covarianze cross-product.** Viene visualizzato per ciascuna coppia di variabili. Il cross-product delle deviazioni 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, uguale alla deviazione cross-product divisa per *N*–1.

Valori mancati. È possibile scegliere tra le opzioni seguenti:

- **Escludi casi a coppie.** 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.
- **Escludi casi a livello di elenco.** 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`).

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

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 ci si potrebbe aspettare che una tale relazione sia di tipo negativo, uno studio riporta una correlazione *positiva* significativa: mentre aumenta il finanziamento per l'assistenza sanitaria, sembrano aumentare i tassi di malattia. 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.

Considerazioni sui dati relativi alle correlazioni parziali

Dati. Utilizzare variabili simmetriche, quantitative.

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

Per ottenere correlazioni parziali

1. Dai menu, scegliere:
   Analizza > **Correlazione** > **Parziale...**
2. Selezionare due o più variabili numeriche per cui è necessario calcolare le correlazioni parziali.
3. Selezionare una o più variabili numeriche di controllo.

Sono inoltre disponibili le seguenti opzioni:

- **Test di significatività.** È possibile selezionare le probabilità a due code o a una coda. Se si conosce in anticipo la direzione dell'associazione, selezionare **A una coda.**. In alternativa, selezionare **A due code**.
- **Visualizza livello di significatività effettivo.** 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

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 mancati. È possibile scegliere una delle seguenti opzioni:

- **Escludi casi a livello di elenco.** I casi con valori mancanti per qualsiasi variabile, incluse le variabili di controllo, vengono esclusi da tutti i calcoli.
- **Escludi casi a coppie.** 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. La cancellazione a coppie consente di utilizzare il massimo numero di dati possibile. Il numero di casi, tuttavia, può differire a seconda del coefficiente. Quando è attiva la cancellazione a coppie, 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.

Funzioni 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`).

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

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 della distanza o similarità possono poi essere utilizzate insieme ad altre procedure quali analisi fattoriale, analisi cluster o scaling multidimensionale per l'analisi di dataset complessi.

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 al quadrato, 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 al quadrato, 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

1. Dai menu, scegliere:

   Analizza > **Correlazione** > **Distanze...**
2. Selezionare una variabile numerica per calcolare la distanza tra casi e almeno due variabili numeriche per calcolare la distanza tra variabili.
3. Scegliere un'alternativa nel gruppo Calcola distanze per calcolare le prossimità tra casi o tra variabili.

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 di intervallo.** Distanza euclidea, distanza euclidea al quadrato, Chebychev, City-block, Minkowski o personalizzata.
- **Dati di conteggio.** Misura chi-quadrato e misura phi-quadrato.
- **Dati binari.** Distanza euclidea, distanza euclidea al quadrato, 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 dei dati per casi o valori *prima* di calcolare le prossimità. 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, grandezza 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 della distanza. Le opzioni disponibili sono Valori assoluti, Cambia segno e Modifica scala su intervallo 0-1.

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 di 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 dei dati per casi o valori prima di calcolare le prossimità. 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, grandezza 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 della distanza. Le opzioni disponibili sono Valori assoluti, Cambia segno e Modifica scala su intervallo 0-1.

Funzioni aggiuntive del comando PROXIMITIES

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

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

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

Capitolo 15. Modelli lineari

I modelli lineari prevedono una destinazione continua in base alle relazioni lineari tra la destinazione e uno o più predittori.

I modelli lineari sono relativamente semplici e offrono una formula matematica di facile interpretazione per il calcolo del punteggio. Le proprietà di questi modelli sono ben note; in genere possono essere creati molto rapidamente rispetto agli altri tipi di modelli (ad esempio le reti neurali o le strutture ad albero delle decisioni) nello stesso dataset.

Esempio. Una compagnia di assicurazioni con risorse limitate per indagare sulle richieste di indennizzo di proprietari di case desidera creare un modello per la stima dei costi delle richieste. Distribuendo questo modello ai centri di servizi, i rappresentanti possono immettere le informazioni sulle richieste di indennizzo mentre sono al telefono con un cliente e ottenere immediatamente il costo "previsto" della richiesta in base ai dati passati. Per ulteriori informazioni, consultare l'argomento .

Requisiti dei campi. Deve esistere una Destinazione e almeno un input. Per impostazione predefinita, i campi con i ruoli predefiniti Entrambi e Nessuno non vengono utilizzati. La destinazione deve essere continua (scala). Non vi sono limitazioni del livello di misurazione sui predittori (input); i campi categoriali (nominale e ordinale) vengono utilizzati come fattori nel modello e i campi continui vengono utilizzati come covariate.

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

Come ottenere un modello lineare

Questa funzione richiede il modulo Statistics Base.

Dai menu, scegliere:

Analizza > **Regressione** > **Modellazione lineare automatica...**

1. Verificare che vi sia almeno una destinazione e un input.
2. Fare clic su **Opzioni di creazione** per specificare le impostazioni facoltative del modello e di creazione.
3. Fare clic su **Opzioni modello** per salvare i punteggi nel dataset attivo ed esportare il modello in un file esterno.
4. Fare clic su **Esegui** per eseguire la procedura e creare gli oggetti Modello.

Obiettivi

Qual è l'obiettivo principale? Selezionare l'obiettivo appropriato.

- **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 agli insiemi di cui è stato eseguito il boosting, il bagging o agli insiemi di dataset di grandi dimensioni.
- **Migliora la precisione del modello (boosting).** Il metodo crea un modello di insieme tramite boosting, che genera una sequenza di modelli per ottenere previsioni più precise. La creazione e il calcolo del punteggio degli insiemi può richiedere più tempo rispetto ai modelli standard.

  Il boosting produce una successione di "modelli di componente", ognuno dei quali viene costruito sul dataset intero. Prima di costruire ogni modello di componente successivo, i record vengono pesati sulla

base dei residui del modello di componente precedente. Ai casi con residui grandi vengono assegnati pesi dell'analisi relativamente più elevati, in modo che il modello del componente successivo si concentrerà sulla previsione corretta di tali record. Insieme questi modelli del componente formano un modello dell'insieme. Il modello dell'insieme calcola il punteggio di nuovi record utilizzando una regola di combinazione; le regole disponibili dipendono dal livello di misurazione della destinazione.

- **Migliora la stabilità del modello (bagging).** Il metodo crea un modello di insieme tramite bagging (aggregazione bootstrap), che genera più modelli per ottenere previsioni più affidabili. La creazione e il calcolo del punteggio degli insiemi può richiedere più tempo rispetto ai modelli standard.

  L'aggregazione di bootstrap (bagging) genera delle repliche del dataset di training, eseguendo il campionamento con sostituzione dal dataset originale. Ciò crea dei campioni di bootstrap di uguale dimensione nel dataset originale. Quindi, un "modello del componente" viene creato in ciascuna replica. Insieme questi modelli del componente formano un modello dell'insieme. Il modello dell'insieme calcola il punteggio di nuovi record utilizzando una regola di combinazione; le regole disponibili dipendono dal livello di misurazione della destinazione.
- **Crea un modello per dataset di grandi dimensioni (è richiesto IBM SPSS Statistics Server).** Il metodo crea un modello di insieme suddividendo il dataset in blocchi di dati. Scegliere questa opzione se il dataset 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 IBM SPSS Statistics Server.

Consultare “Insiemi” a pagina 63 per le impostazioni relative a boosting, bagging e ai dataset molto grandi.

Di base

Prepara dati automaticamente. Questa opzione consente alla procedura di trasformare internamente la destinazione e i predittori per espandere al massimo la potenza predittiva del modello; le 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 impostazione predefinita, viene eseguita le preparazioni automatica dei dati riportata di seguito.

- **Gestione di data e ora.** Ciascun predittore della data viene trasformato in un nuovo predittore continuo contenente il tempo trascorso da una data di riferimento (1970-01-01). Ogni predittore dell'ora viene trasformato in un nuovo predittore continuo contenente l'ora trascorsa da un'ora di riferimento (00:00:00).
- **Adatta 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 trovano oltre un valore di interruzione (3 deviazioni standard dalla media) vengono impostati sul valore di interruzione.
- **Gestione valori mancanti.** I valori mancanti dei predittori nominali vengono sostituiti con la modalità della partizione di training. I valori mancanti dei predittori ordinali vengono sostituiti dalla mediana della partizione di training. I valori mancanti dei predittori continui vengono sostituiti dalla media della partizione di training.
- **Unione supervisionata.** Ciò rende più parsimonioso un modello, riducendo il numero di campi da elaborare insieme alla destinazione. Le categorie simili vengono identificate in base alla relazione tra input e obiettivo. Le categorie che non sono significativamente differenti (ovvero, che hanno un valore P maggiore di 0.1) vengono unite. Se tutte le categorie vengono unite in un'unica categoria, le versioni originali e derivate del campo vengono escluse dal modello perché non hanno alcun valore come predittore.

Livello di confidenza. Si tratta del livello di confidenza utilizzato per calcolare le stime dell'intervallo dei coefficienti del modello nella vista Coefficienti. Specificare un valore maggiore di 0 e minore di 100. Il valore predefinito è 95.

Selezione modello

Metodo di selezione modello. Scegliere uno dei metodi di selezione del modello (dettagli di seguito) o **Includi tutti i predittori**, che semplicemente immette tutti i predittori disponibili come termini del modello di effetti principali. Per impostazione predefinita. viene utilizzata l'opzione **A fasi in avanti**.

Selezione a fasi in avanti. All'inizio non vi sono effetti nel modello e questi vengono aggiunti e rimossi una fase per volta finché non possono più essere aggiunti o rimossi in base ai criteri a fasi.

- **Criteri per immissione/rimozione.** Si tratta della statistica utilizzata per determinare se un effetto deve essere aggiunto o rimosso nel modello. **Criterio di informazione (AICC)** si basa sulla verosimiglianza dell'insieme di training, dato il modello, e viene adattato per penalizzare i modelli eccessivamente complessi. **Statistiche F** si basano su un test statistico del miglioramento nell'errore del modello. **R-quadrato adattato** si basa sull'adattamento dell'insieme di training e viene adattato per penalizzare i modelli eccessivamente complessi. **Criterio di prevenzione del sovradattamento (ASE)** si basa sull'adattamento (ASE, average squared error) dell'insieme di prevenzione sovradattato. L'insieme di prevenzione sovradattato è un sottocampione random di circa il 30% del dataset originale che non è utilizzato per formare il modello.

  Se si sceglie un criterio diverso da **Statistiche F**, in ogni fase l'effetto corrispondente all'incremento positivo più grande nel criterio viene aggiunto al modello. Gli effetti nel modello che corrispondono a un decremento nel criterio vengono rimossi.

  Se si sceglie **Statistiche F** come criterio, in ogni fase l'effetto che ha il valore *p* più piccolo inferiore alla soglia specificata, **Includi effetti con valori p inferiori a**, viene aggiunto al modello. Il valore predefinito è 0.05. Gli effetti nel modello con un valore *p* superiore alla soglia specificata, **Rimuovi effetti con valori p maggiori di**, vengono rimossi. Il valore predefinito è 0.10.
- **Personalizza numero massimo di effetti nel modello finale.** Per impostazione predefinita, tutti gli effetti disponibili possono essere immessi nel modello. In alternativa, se l'algoritmo a fasi termina una fase con il numero massimo di effetti specificato, l'algoritmo termina con l'insieme corrente di effetti.
- **Personalizza numero massimo di fasi.** L'algoritmo a fasi termina dopo un determinato numero di fasi. Per impostazione predefinita, è tre volte il numero di effetti disponibili. In alternativa, specificare un numero intero positivo come numero massimo di fasi.

Selezione sottoinsiemi migliori. Controlla "tutti i possibili" modelli o almeno un sottoinsieme di modelli possibili più grande di quello a fasi in avanti per scegliere il migliore in base al criterio Sottoinsiemi migliori. **Criterio di informazione (AICC)** si basa sulla verosimiglianza dell'insieme di training, dato il modello, e viene adattato per penalizzare i modelli eccessivamente complessi. **R-quadrato adattato** si basa sull'adattamento dell'insieme di training e viene adattato per penalizzare i modelli eccessivamente complessi. **Criterio di prevenzione del sovradattamento (ASE)** si basa sull'adattamento (ASE, average squared error) dell'insieme di prevenzione sovradattato. L'insieme di prevenzione sovradattato è un sottocampione random di circa il 30% del dataset originale che non è utilizzato per formare il modello.

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

Nota: la selezione dei sottoinsiemi migliori richiede attività di calcolo più intense rispetto alla selezione a fasi in avanti. Quando si utilizza la selezione Sottoinsiemi migliori in combinazione con attività di boosting, bagging o di dataset molto grandi, la creazione può richiedere molto più tempo di un modello standard creato utilizzando la selezione a fasi in avanti.

Insiemi

Queste impostazioni determinano il funzionamento della procedura di insieme che si verifica in caso di richiesta di boosting, bagging o dataset di grandi dimensioni negli obiettivi. Le opzioni che non si applicano all'obiettivo selezionato vengono ignorate.

Bagging e dataset di grandi dimensioni. Quando si calcola il punteggio di un insieme, questa è la regola utilizzata per combinare i valori previsti dai modelli di base per calcolare il valore del punteggio dell'insieme.

- **Regola di combinazione predefinita per le destinazioni continue.** I valori previsti dell'insieme per le destinazioni continue possono essere combinati utilizzando la media o la mediana dei valori previsti ricavati dai modelli di base.

Si osservi che quando l'obiettivo consiste nel migliorare la precisione del modello, le selezioni delle regole di combinazione vengono ignorate. Il boosting utilizza sempre un voto di maggioranza pesato per calcolare il punteggio delle destinazioni categoriali e una mediana pesata per calcolare il punteggio delle destinazioni continue.

Boosting e bagging. Specifica il numero di modelli di base da creare quando un obiettivo consiste nel migliorare la precisione o la stabilità del modello; per il bagging, si tratta del numero di campioni di bootstrap. Deve essere un numero intero positivo.

Avanzate

Replica risultati. L'impostazione di un seme random consente di replicare le analisi. Il generatore di numeri random viene utilizzato per scegliere i record presenti nell'insieme di prevenzione del sovradattamento. Specificare un numero intero o fare clic su **Genera** per creare un numero intero pseudo-random compreso tra 1 e 2147483647, compresi. Il valore predefinito è 54752075.

Opzioni modello

Salva valori previsti nel dataset. Il nome predefinito della variabile è *PredictedValue*.

Esporta modello. Scrive il modello in un file *.zip* esterno. È possibile utilizzare questo file 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 specifica del file si riferisce ad un file esistente, il file viene sovrascritto.

Riepilogo del modello

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

Tabella. La tabella identifica alcune impostazioni del modello di alto livello, tra cui:

- Il nome della destinazione specificata 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. Inoltre viene visualizzato il valore del criterio di selezione per il modello finale, che viene presentato in un formato che predilige le dimensioni più piccole.

Grafico. Il grafico visualizza l'accuratezza del modello finale, che viene presentato in un formato che predilige le dimensioni più grandi. Il valore è $100 \times R^2$ adattato per il modello finale.

Preparazione automatica dati

Questa vista mostra le informazioni su quali campi sono esclusi e su come vengono derivati i campi trasformati nel passo ADP (automatic data preparation, preparazione automatica dati). Per ogni campo che è stato trasformato o escluso, la tabella elenca il nome del campo, il suo ruolo nell'analisi e l'azione eseguita dal passo ADP. I campi sono disposti in ordine alfabetico crescente dei nomi dei campi. Le azioni che è possibile eseguire su ciascun campo includono:

- **Deriva durata: mesi** calcola il tempo trascorso, in mesi, dai valori in un campo contenente le date fino alla data corrente del sistema.

- **Deriva durata: ore** calcola il tempo trascorso, in ore, dai valori in un campo contenente le ore fino alla data corrente del sistema.
- **Modifica livello di misurazione da continuo a ordinale** ricompone i campi continui con meno di 5 valori univoci come campi ordinali.
- **Modifica livello di misurazione da ordinale a continuo** ricompone i campi ordinali con più di 10 valori univoci come campi continui.
- **Ritaglia valori anomali** imposta i valori dei predittori continui che si trovano oltre un valore di interruzione (3 deviazioni standard dalla media) per il valore di interruzione.
- **Sostituisci valori mancanti** sostituisce i valori mancanti dei campi nominali con la modalità, dei campi ordinali con la mediana e dei campi continui con la media.
- **Unisci categorie per aumentare al massimo l'associazione alla destinazione** identifica le categorie di predittori "simili" in base alla relazione tra l'input e la destinazione. Le categorie che non sono significativamente differenti (ovvero, che hanno un valore *p* maggiore di 0.05) vengono unite.
- **Escludi predittore costante / dopo gestione dei valori anomali / dopo l'unione di categorie** rimuove i predittori che hanno un solo valore, probabilmente dopo che sono state intraprese altre azioni ADP.

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

Visualizza un grafico a dispersione raccolto dei valori previsti sull'asse verticale in base ai valori osservati sull'asse orizzontale. Idealmente, i punti devono giacere su una linea di 45 gradi; questa vista può indicare eventuali record previsti in modo particolarmente non corretto dal modello.

Residui

Visualizza un grafico diagnostico dei residui del modello.

Stili del grafico. Sono disponibili diversi stili di visualizzazione, che sono accessibili dall'elenco a discesa **Stile**.

- **Istogramma.** Si tratta di un istogramma raccolto dei residui studentizzati con una sovrapposizione della distribuzione normale. I modelli lineari presumono che i residui abbiano una distribuzione normale, quindi l'istogramma idealmente deve avvicinarsi molto alla linea uniforme (piatta).
- **Grafico P-P.** Si tratta di un grafico probabilità-probabilità raccolto che confronta i residui studentizzati con una distribuzione normale. Se la pendenza dei punti tracciati è meno ripida della linea normale, i residui mostrano più variabilità di una distribuzione normale; se la pendenza è più ripida, i residui mostrano meno variabilità di una distribuzione normale. Se i punti tracciati hanno una curva a forma di S, la distribuzione dei residui è asimmetrica.

Valori anomali

Questa tabella elenca i record che esercitano un'influenza non necessaria sul modello e visualizza l'ID record (se specificato nella scheda Campi), il valore di destinazione e la distanza di Cook. La distanza di Cook è una misura di quando i residui di tutti i record cambiano se un determinato record è stato escluso dal calcolo dei coefficienti del modello. Una distanza di Cook di grandi dimensioni indica che l'esclusione di un record dal calcolo cambia notevolmente i coefficienti, pertanto deve essere considerata influente.

I record influenti devono essere esaminati con attenzione per determinare se è possibile assegnare ad essi meno peso nella stima del modello o troncare i valori anomali ad una soglia accettabile oppure rimuovere completamente i record influenti.

Effetti

Questa vista mostra la dimensione di ciascun effetto nel modello.

Stili. Sono disponibili diversi stili di visualizzazione, che sono accessibili dall'elenco a discesa **Stile**.

- **Diagramma.** Si tratta di un grafico in cui gli effetti vengono ordinati dall'alto verso il basso, riducendo l'importanza del predittore. Le linee di connessione nel diagramma sono pesate in base alla significatività dell'effetto, per cui la larghezza della linea più grande corrisponde agli effetti più significativi (valori *p* più piccoli). Passando con il cursore del mouse su una linea di connessione viene visualizzato un suggerimento che mostra il valore *p* e l'importanza dell'effetto. È l'impostazione predefinita.
- **Tabella.** Si tratta di una tabella ANOVA per gli effetti del modello globale e del modello individuale. Gli effetti individuali vengono ordinati dall'alto verso il basso, riducendo l'importanza del predittore. Si osservi che per impostazione predefinita la tabella è compressa per mostrare solo i risultati relativi al modello globale. Per visualizzare i risultati relativi agli effetti del modello individuale, fare clic sulla cella **Modello corretto** nella tabella.

Importanza predittore. Esiste un dispositivo di scorrimento Importanza predittore che controlla quali predittori vengono visualizzati nella vista. Questo non cambia il modello, ma consente semplicemente di concentrare l'attenzione sui predittori più importanti. Per impostazione predefinita, vengono visualizzati i primi 10 effetti.

Significatività. Esiste un dispositivo di scorrimento Significatività che controlla ulteriormente quali effetti vengono mostrati nella vista oltre a quelli mostrati in base all'importanza del predittore. Gli effetti con valori di significatività superiori al valore del dispositivo di scorrimento vengono nascosti. Questo non cambia il modello, ma consente semplicemente di concentrare l'attenzione sugli effetti più importanti. Per impostazione predefinita, il valore è 1.00, quindi non viene filtrato alcun effetto in base alla significatività.

Coefficienti

Questa vista mostra il valore di ciascun coefficiente nel modello. Si osservi che i fattori (predittori categoriali) sono indicatori codificati nel modello, in tal modo gli **effetti** contenenti i fattori avranno, in genere, più **coefficienti** associati; uno per ogni categoria, ad eccezione della categoria corrispondente al parametro ridondante (riferimento).

Stili. Sono disponibili diversi stili di visualizzazione, che sono accessibili dall'elenco a discesa **Stile**.

- **Diagramma.** Si tratta di un grafico che visualizza prima l'intercettazione e poi ordina gli effetti dall'alto verso il basso, riducendo l'importanza del predittore. All'interno degli effetti che contengono fattori, i coefficienti sono disposti in ordine crescente di valori dei dati. Le linee di connessione nel diagramma sono colorate in base al segno del coefficiente (vedere la chiave del diagramma) e pesate in base alla significatività del coefficiente; la larghezza della linea più grande corrisponde ai coefficienti più significativi (valori *p* più piccoli). Passando con il cursore del mouse su una linea di connessione viene visualizzato un suggerimento 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 predefinito.
- **Tabella.** Mostra i valori, i test di significatività e gli intervalli di confidenza per i singoli coefficienti del modello. Dopo l'intercettazione, gli effetti vengono ordinati dall'alto verso il basso, riducendo l'importanza del predittore. All'interno degli effetti che contengono fattori, i coefficienti sono disposti in ordine crescente di valori dei dati. Si osservi che per impostazione predefinita la tabella è compressa per mostrare solo il coefficiente, la significatività e l'importanza di ciascun parametro del modello. Per visualizzare l'errore standard, la statistica *t* e l'intervallo di confidenza, fare clic sulla cella **Coefficiente** nella tabella. Passando con il cursore del mouse sul nome di un parametro del modello nella tabella,

viene visualizzato un suggerimento che mostra il nome del parametro, l'effetto a cui è associato il parametro e (per i predittori categoriali) le etichette dei valori associate 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. Esiste un dispositivo di scorrimento Importanza predittore che controlla quali predittori vengono visualizzati nella vista. Questo non cambia il modello, ma consente semplicemente di concentrare l'attenzione sui predittori più importanti. Per impostazione predefinita, vengono visualizzati i primi 10 effetti.

Significatività. Esiste un dispositivo di scorrimento Significatività che controlla ulteriormente quali coefficienti vengono mostrati nella vista oltre a quelli mostrati in base all'importanza del predittore. I coefficienti con valori di significatività superiori al valore del dispositivo di scorrimento vengono nascosti. Questo non cambia il modello, ma consente semplicemente di concentrare l'attenzione sui coefficienti più importanti. Per impostazione predefinita, il valore è 1.00, quindi non viene filtrato alcun coefficiente in base alla significatività.

Medie stimate

Si tratta dei grafici visualizzati per i predittori significativi. Il grafico visualizza il valore stimato dal modello della destinazione sull'asse verticale per ciascun valore del predittore sull'asse orizzontale, mantenendo costanti tutti gli altri predittori. Offre una visualizzazione utile degli effetti dei coefficienti di ciascun predittore sulla destinazione.

Nota: se nessuno dei predittori è significativo, non viene generata alcuna media stimata.

Riepilogo creazione modello

Quando si sceglie un algoritmo di selezione del modello diverso da **Nessuno** nelle impostazioni di Selezione modello, vengono forniti alcuni dettagli del processo di creazione del modello.

A fasi in avanti. Quando A fasi in avanti è l'algoritmo di selezione, la tabella visualizza le ultime 10 fasi nell'algoritmo A fasi in avanti. Per ogni fase vengono mostrati il valore del criterio di selezione e gli effetti nel modello nella fase in questione. In questo modo, si ha un'idea del contributo dato da ciascuna fase nel modello. Ogni colonna consente di ordinare le righe, in modo che sia più facile vedere quali effetti sono presenti nel modello in una determinata fase.

Sottoinsiemi migliori. Quando Sottoinsiemi migliori è l'algoritmo di selezione, la tabella visualizza i primi 10 modelli. Per ogni modello vengono mostrati il valore del criterio di selezione e gli effetti nel modello. In tal modo, si ha un'idea della stabilità dei primi modelli; se tendono ad avere molti effetti simili con poche differenze, si può essere abbastanza fiduciosi sul "primo" modello; se tendono ad avere effetti molto diversi, alcuni degli effetti potrebbero essere troppo simili e devono essere combinati (o uno di essi rimosso). Ogni colonna consente di ordinare le righe, in modo che sia più facile vedere quali effetti sono presenti nel modello in una determinata fase.

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 multiplo, R^2, R^2 adattato, modifica in R^2, errore standard della stima, tabella di analisi della varianza, valori previsti e residui. Inoltre, intervalli di confidenza del 95% per ogni coefficiente di regressione, matrice della varianza-covarianza, fattore di inflazione della varianza, tolleranza, test di Durbin-Watson, misure della distanza (Mahalanobis, Cook e valori di leva), DfBeta, DfFit, intervalli di previsione e informazioni di diagnostica per casi. Grafici: grafici a dispersione, grafici parziali, istogrammi e grafici di probabilità normale.

Considerazioni sui dati della regressione lineare

Dati. Le variabili dipententi e 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 (dummy) o altri tipi di variabili di contrasto.

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 ogni variabile indipendente deve essere lineare e tutte le osservazioni devono essere indipendenti.

Per ottenere un'analisi della regressione lineare

1. Dai menu, scegliere:

   Analizza > **Regressione** > **Lineare...**
2. Nella finestra di dialogo Regressione lineare, selezionare una variabile dipendente numerica.
3. 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 peso per WLS per un'analisi dei minimi quadrati pesati.

Minimi quadrati pesati (WLS). Consente di ottenere un modello dei minimi quadrati pesati. I punti dati vengono pesati in base al reciproco della loro varianza. Le osservazioni con varianza elevata hanno un

peso minore nell'analisi rispetto a quelle con varianza ridotta. Se il valore della variabile peso è zero, negativo o mancante, il caso viene escluso 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.

- *Inserimento (regressione)*. Una procedura per la selezione delle variabili in cui tutte le variabili in un blocco sono inserite in una singola fase.
- *A fasi*. Ad ogni fase 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 nessun'altra variabile rispetta il criterio di inserimento o quello di rimozione.
- *Rimuovi*. Una procedura per la selezione delle variabili in cui tutte le variabili in un blocco sono rimosse in una singola fase.
- *Eliminazione all'indietro*. Una procedura di selezione di variabili nella quale tutte le variabili vengono inserite nell'equazione e quindi 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 a fasi 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 a fasi (a fasi, 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 a fasi 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

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

I grafici possono facilitare la convalida 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, nell'Editor dei dati sono disponibili i valori previsti, i residui e altre informazioni di diagnostica per creare dei grafici con le variabili indipendenti. Sono disponibili i seguenti grafici:

Grafici a dispersione. È possibile tracciare due qualsiasi dei seguenti grafici: variabile dipendente, valori previsti standardizzati, residui standardizzati, residui eliminati, valori previsti adattati, residui studentizzati o residui eliminati studentizzati. Rappresentare nel grafico i residui standardizzati e i valori attesi standardizzati per verificare la linearità e l'uguaglianza delle varianze.

Elenco di variabili origine. Elenca la variabile dipendente (DEPENDNT) e le seguenti variabili previste e residue: 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 . È possibile ottenere gli istogrammi dei residui standardizzati e i grafici di probabilità normale confrontando la distribuzione dei residui standardizzati con una distribuzione normale.

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

Regressione lineare: salvataggio di nuove variabili

È possibile salvare i valori previsti, i residui e altre statistiche utili per le informazioni di diagnostica. Ogni selezione aggiunge una o più nuove variabili al file di dati attivo.

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

- *Non standardizzati*. Il valore previsto dal modello per la variabile dipendente.
- *Standardizzati*. Una trasformazione di ciascun valore previsto nella sua forma standardizzata. Ossia, il valore previsto medio viene sottratto dal valore previsto e la differenza viene divisa per la deviazione standard dei valori previsti. I valori previsti standardizzati hanno una media pari a 0 e una deviazione standard pari a 1.
- *Adattati*.. Il valore previsto per un caso quando tale caso viene escluso dal calcolo dei coefficienti di regressione.
- *Errore standard delle previsioni delle medie*. Errori standard dei valori previsti. 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 di quanto i valori di un caso sulle variabili indipendenti differiscono dalla media di tutti i casi. Un'elevata distanza di Mahalanobis identifica un caso come caratterizzato da 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. Un valore alto del D di Cook indica che l'esclusione di un caso dal calcolo delle statistiche di regressione modifica sostanzialmente i coefficienti.
- *Valori di leva*. Misura l'influenza di un punto sull'adattamento della regressione. La leva centrata è compresa tra 0 (nessuna influenza sull'adattamento) e (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 previsione della risposta prevista media.
- *Singolo*. Limiti inferiore e superiore (due variabili) per l'intervallo di previsione della variabile dipendente per un singolo caso.
- *Intervallo di confidenza*. Immettere un valore compreso tra 1 e 99,99 per specificare il livello di confidenza per i due intervalli di previsione. Per disporre di questa opzione è necessario aver selezionato Media o Individuale. I valori di intervallo di confidenza tipici 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 previsto dal modello.
- *Standardizzati*. Il residuo diviso per una stima della sua deviazione standard. I residui standardizzati, noti anche come residui di Pearson, hanno una media pari a 0 e una deviazione standard pari a 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 ciascun caso nelle variabili indipendenti e le medie delle variabili indipendenti.
- *Eliminato*. Il residuo per un caso quando tale caso viene escluso dal calcolo dei coefficienti di regressione. È la differenza tra il valore della variabile dipendente e il valore previsto adattato.
- *Per cancellazione studentizzati*. Il residuo cancellato per un caso diviso per il proprio 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.

- *DiffBeta*. La differenza nel valore beta nella variazione nel coefficiente di regressione risultante dall'esclusione di uno specifico caso. Un valore viene calcolato per ogni termine nel modello, compreso quello costante.
- *Differenza standardizzata in beta*. 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 di casi. Un valore viene calcolato per ogni termine nel modello, compreso quello costante.
- *DifAdattamento*. La differenza nel valore di adattamento è la variazione nel valore previsto risultante dall'esclusione di uno specifico caso.
- *Differenza standardizzata nell'adattamento*. La differenza standardizzata nel valore di adattamento. 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.*. Il rapporto tra il determinante della matrice di covarianza con uno specifico caso escluso dal calcolo dei coefficienti di regressione e il determinante della matrice di covarianza con tutti i casi inclusi. Se il rapporto è prossimo a 1, il caso non modifica in modo significativo la matrice di covarianza.

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

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

Regressione lineare: Statistiche

Sono disponibili le seguenti statistiche:

Coefficienti di regressione. Stime visualizza il coefficiente di regressione *B*, l'errore standard di *B*, il valore beta del coefficiente standardizzato, il valore *t* per *B* e il livello di significatività di *t* a due code. **Intervalli di confidenza** visualizza gli intervalli di confidenza con il livello specificato di confidenza per ogni coefficiente di regressione o una 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 immesse e rimosse dal modello e vengono visualizzate le seguenti statistiche della bontà di adattamento: *R* multiplo, R^2 e R^2 adattato, errore standard della stima e una tabella dell'analisi della varianza.

Cambiamento di R quadrato. Il cambiamento nella statistiche R^2 che viene prodotto aggiungendo o eliminando una variabile indipendente. Se la variazione di R^2 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. La correlazione residua fra due variabili, dopo aver rimosso la correlazione dovuta alla loro reciproca associazione alle altre variabili. La correlazione fra la variabile dipendente e una variabile indipendente dopo aver rimosso da entrambe gli effetti lineari delle altre variabili indipendenti nel modello.

Correlazione parziale. La correlazione fra la variabile dipendente e una variabile indipendente dopo che sono stati rimossi dalla variabile indipendente gli effetti lineari delle altre variabili indipendenti nel modello. 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 cross-product 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. Visualizza il test di Durbin-Watson per la correlazione seriale dei residui e le informazioni di diagnostica per casi per i casi che soddisfano il criterio di selezione (valori anomali oltre *n* deviazioni standard).

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 a fasi. 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*. Una 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 Per rimuovere un numero maggiore di variabili dal modello, ridurre il valore di rimozione.
- *Usa valore 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 Per rimuovere un numero maggiore di variabili dal modello, aumentare il valore di rimozione.

Includi 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^2 non può essere interpretato nel modo usuale.

Valori mancati. È possibile scegliere tra le opzioni seguenti:

- **Escludi casi a livello di elenco.** Sono inclusi nell'analisi solo i casi con valori validi per tutte le variabili.
- **Escludi casi a coppie.** 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* minimo a coppie.
- **Sostituisci con la media.** Per i calcoli vengono utilizzati tutti i casi e la media della variabile viene sostituita alle osservazioni mancanti.

Funzioni 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 progettazione 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 delle differenze al quadrato tra una variabile di risposta (dipendente) e una combinazione pesata 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 cumulative, residui di Pearson per le frequenze e frequenze cumulative, probabilità osservate e attese, probabilità osservate e probabilità cumulative attese per ciascuna categoria di risposta in base al modello di covariata, correlazione asintotica e matrici di covarianza delle stime del parametro, chi-quadrato di Pearson e chi-quadrato del rapporto di verosimiglianza, statistiche della bontà di adattamento, cronologia delle iterazioni, test sull'ipotesi di linee parallele, stime del parametro, errori standard, intervalli di confidenza e statistiche R^2 di Cox e Snell, di Nagelkerke e di McFadden.

Considerazioni sui dati della procedura Regressione ordinale

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.

Ipotesi. 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.

Come ottenere una regressione ordinale

1. Dai menu, scegliere:
   Analizza > **Regressione** > **Ordinale...**
2. Selezionare una variabile dipendente.
3. Fare clic su **OK**.

Regressione ordinale: Opzioni

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

Iterazioni. È possibile personalizzare l'algoritmo iterativo.

- **Numero massimo di iterazioni.** Specificare un intero non negativo. Se si specifica 0, la procedura restituisce le stime iniziali.
- **Max dimezzamenti.** Specificare un intero positivo.
- **Convergenza verosimiglianza logaritmica.** L'algoritmo si interrompe se il cambiamento assoluto o relativo nella verosimiglianza logaritmica 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 del parametro 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 Collegamento. La funzione Collegamento è la trasformazione delle probabilità cumulative che permette di stimare il modello. Sono disponibili le cinque funzioni Collegamento riportate di seguito.

- **Logit.** $f(x)=\log(x/(1-x))$. In genere, viene utilizzata per le categorie distribuite in modo uniforme.
- **Log-log complementare.** $f(x)=\log(-\log(1-x))$. In genere, viene utilizzata quando le categorie più alte sono più probabili.
- **Log-log negativa.** $f(x)=-\log(-\log(x))$. In genere, viene utilizzata quando le categorie più basse sono più probabili.
- **Probit.** $f(x)=\Phi^{-1}(x)$. In genere, viene utilizzata quando la variabile latente viene distribuita normalmente.
- **Cauchit (Cauchy inverso).** $f(x)=\tan(\pi(x-0.5))$. In genere, viene utilizzata quando la variabile latente ha molti valori estremi.

Regressione ordinale: Output

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

Visualizza. Consente di creare tabelle per:

- **Stampa cronologia delle iterazioni.** Stampa le stime del parametro e della verosimiglianza logaritmica per la frequenza di iterazioni di stampa specificata. La prima e l'ultima iterazione vengono sempre stampate.
- **Statistiche della bontà di adattamento.** Statistica chi-quadrato di Pearson e dei rapporti di verosimiglianza. Vengono elaborate in base alla classificazione specificata nell'elenco di variabili.
- **Statistiche di riepilogo.** Statistiche R^2 di Cox e Snell, di Nagelkerke e di McFadden.
- **Stime dei parametri.** Stime del parametro, errori standard e intervalli di confidenza.
- **Correlazione asintotica delle stime del parametro.** Matrice delle correlazioni delle stime dei parametri.
- **Covarianza asintotica delle stime del parametro.** Matrice delle covarianze delle stime dei parametri.
- **Informazioni sulle celle.** Frequenze osservate e previste e frequenze cumulative, residui di Pearson per frequenze e frequenze cumulative, probabilità osservate e previste, probabilità cumulative osservate e

previste per ciascuna categoria di risposta in base al modello di covariata. Si noti che per i modelli che includono più modelli di covariate, 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 dell'ipotesi di equivalenza dei parametri di ubicazione nei livelli della variabile dipendente. È disponibile per il modello di sola ubicazione.

Variabili salvate. Salva le seguenti variabili nel file di lavoro:
- **Probabilità di risposta stimate.** Probabilità stimate dal modello per la classificazione di un modello di fattore o covariata 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 di fattore o covariata.
- **Probabilità di categoria prevista.** Probabilità stimata di classificazione di un modello di fattore o covariata nella categoria prevista. La probabilità corrisponde inoltre al massimo di probabilità stimate del modello di fattore o covariata.
- **Probabilità di categoria reale.** Probabilità stimata di classificazione di un modello di fattore o covariata nella categoria reale.

Stampa verosimiglianza logaritmica. Controlla la visualizzazione della verosimiglianza logaritmica. 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: Ubicazione

Nella finestra di dialogo Ubicazione è possibile specificare il modello di ubicazione per l'analisi.

Specifica modello. Un modello di 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 dei fattori o di interazioni di covariate.

Fattori/covariate. Vengono elencati i fattori e le covariate.

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

Crea 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.

Fattori/covariate. Vengono elencati i fattori e le covariate.

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

Crea 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.

Funzioni aggiuntive del comando PLUM

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

- Creare test sull'ipotesi personalizzati specificando ipotesi null come combinazioni lineari dei parametri.

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

Capitolo 18. Curva stimata

La procedura Curva stimata produce le statistiche di regressione per la Curva stimata e i grafici correlati per 11 diversi modelli di regressione per la curva stimata. 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 previsione.

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 ipotesi e la bontà di adattamento del modello.

Statistiche. Per ogni modello: coefficienti di regressione, *R* multiplo, R^2, R^2 adattato, errore standard della stima, tabella di analisi della varianza, valori previsti, residui e intervalli di previsione. Modelli: lineare, logaritmico, inverso, quadratico, cubico, potenza, composto, curva S, logistico, crescita ed esponenziale.

Considerazioni sui dati relativi alla curva stimata

Dati. Le variabili dipendenti ed indipendenti devono essere quantitative. Se si seleziona **Tempo** come variabile indipenente dal dataset attivo anziché selezionare una variabile, la procedura Curva stimata 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.

Ipotesi. Rappresentare i dati graficamente per determinare come sono correlate le variabili indipendenti e dipendenti (in modo lineare, esponenziale e così via). 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 ogni 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 curva stimata

1. Dai menu, scegliere:

   Analizza > **Regressione** > **Curva stimata...**
2. Selezionare una o più variabili dipendenti. Per ciascuna variabile dipendente verrà creato un modello distinto.
3. Selezionare una variabile indipendente (una variabile nel dataset attivo oppure **Tempo**).
4. 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 previsione come nuove variabili.

Sono inoltre disponibili le seguenti opzioni:

- **Includi costante nell'equazione.** Consente di valutare un termine costante nell'equazione di regressione. La costante viene inclusa per impostazione predefinita.

- **Modelli di 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.

Curva stimata: Modelli

È possibile scegliere uno o più modelli di regressione per la curva stimata. 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 modellati 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)) o ln(Y) = b0 + (b1/t).

Logistico. 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 valore di limite superiore. Per usare l'equazione di regressione, specificare il valore limite superiore dopo aver selezionato Logistico. Il valore deve essere un positivo maggiore del valore più alto della variabile dipendente.

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

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

Curva stimata: Salva

Salva Variabili. Per ogni modello selezionato è possibile salvare i valori previsti, i residui (valore osservato della variabile dipendente meno il valore previsto del modello) e gli intervalli di previsione (limiti 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 dataset 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 il termine 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 dei minimi quadrati parziali

La procedura Regressione parziale dei 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 alla modellazione 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, una fase di regressione consente di prevedere i valori delle variabili dipendenti mediante la decomposizione delle variabili indipendenti.

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

Grafici. VIP (Variable importance in projection), punteggi fattore, pesi fattore per i primi tre fattori latenti e distanza per il modello sono tutti prodotti dalla scheda Opzioni.

Considerazioni sui dati della regressione parziale dei minimi quadrati

Livello di misurazione. Le variabili dipendenti e indipendenti (predittore) possono essere di scala, nominali o ordinali. La procedura presume che a tutte le variabili sia stato assegnato il livello di misurazione appropriato; sebbene sia possibile modificare temporaneamente il livello di misurazione per una variabile facendo clic con il tasto destro del mouse sulla variabile nell'elenco di variabili di origine e selezionando un livello di misurazione dal menu a comparsa. 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 dummy; ovvero, semplicemente omettendo l'indicatore corrispondente alla categoria di riferimento.

Peso della frequenza. 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 > **Regressione** > **Minimi quadrati parziali...**

1. Selezionare almeno una variabile dipendente.

2. 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 unico per l'output per i casi e i dataset salvati.
- Specificare un limite superiore per il numero dei fattori latenti da estrarre.

Prerequisiti

La Regressione dei minimi quadrati parziali è un comando di estensione Python e richiede IBM SPSS Statistics - Essentials for Python, che è installato per impostazione predefinita con il prodotto IBM SPSS Statistics. Richiede anche le librerie Python NumPy e SciPy Python, disponibili gratuitamente.

Nota: per gli utenti che lavorano in modalità di analisi distribuita (richiede IBM SPSS Statistics Server), NumPy e SciPy devono essere installati sul server. Contattare l'amministratore del sistema per assistenza.

Utenti Windows e Mac

Per Windows e Mac, NumPy e SciPy devono essere installati in una versione di Python 2.7 separata dalla versione installata con IBM SPSS Statistics. Se non si dispone di una versione separata di Python 2.7, è possibile scaricarla da http://www.python.org. Installare quindi NumPy e SciPy per Python versione 2.7. I programmi di installazione sono disponibili da http://www.scipy.org/Download.

Per abilitare l'utilizzo di NumPy e SciPy, è necessario impostare l'ubicazione di Python sulla versione di Python 2.7 dove sono stati installati NumPy e SciPy. L'ubicazione di Python viene impostata dalla scheda Ubicazioni file nella finestra di dialogo Opzioni (Modifica > Opzioni).

Utenti Linux

Si consiglia di scaricare il codice sorgente e di creare NumPy e SciPy personalmente. Il codice sorgente è disponibile da http://www.scipy.org/Download. È possibile installare NumPy e SciPy alla versione di Python 2.7 installata con IBM SPSS Statistics. Si trova nella directory `Python` nell'ubicazione dove è installato IBM SPSS Statistics.

Se si sceglie di installare NumPy ed SciPy a una versione di Python 2.7 diversa dalla versione che è stata installata con IBM SPSS Statistics, è necessario impostare l'ubicazione di Python per puntare a tale versione. L'ubicazione di Python viene impostata dalla scheda Ubicazioni file nella finestra di dialogo Opzioni (Modifica > Opzioni).

Windows e Unix Server

NumPy e SciPy devono essere installati, sul server, in una versione di Python 2.7 separata dalla versione installata con IBM SPSS Statistics. Se non è presente una versione separata di Python 2.7 sul server, è possibile scaricarla da http://www.python.org. NumPy e SciPy per Python 2.7 sono disponibili da http://www.scipy.org/Download. Per abilitare l'utilizzo di NumPy e SciPy, è necessario impostare l'ubicazione di Python per il server sulla versione di Python 2.7 dove sono installati NumPy e SciPy. L'ubicazione di Python è impostata dalla IBM SPSS Statistics Administration Console.

Modello

Specifica effetti del modello. 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.

Crea 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

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 dataset. I nomi dei dataset devono essere univoci. Se si specifica il nome di un dataset esistente, i suoi contenuti vengono sostituiti; altrimenti, viene creato un nuovo dataset.

- **Salva stime per singoli casi.** Salva le seguenti stime del modello per casi: valori previsti, residui, distanza al modello fattore latente e punteggi fattore latente. Inoltre, rende su grafico i punteggi fattoriali latenti.
- **Salva stime per fattori latenti.** Consente di salvare i caricamenti e i pesi fattoriali latenti. Inoltre, rende su grafico i pesi fattoriali latenti.
- **Salva stime per variabili indipendenti.** Consente di salvare le stime del parametro di regressione e l'importanza delle variabili per la proiezione (VIP). Inoltre, rende su grafico i VIP per fattore latente.

Capitolo 20. Analisi della approssimità

L'analisi della approssimità 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 vicini tra loro vengono detti "vicini." Quando viene presentato un nuovo caso (holdout), viene calcolata la sua distanza da ogni caso del modello. Le classificazioni dei casi più simili, i vicini più vicini, vengono conteggiati e il nuovo caso viene inserito nella categoria contenente il numero maggiore di vicini più vicini.

È possibile specificare il numero degli elementi adiacenti più vicini da esaminare; questo valore viene denominato *k*.

L'analisi della approssimità 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.

Considerazioni sui dati dell'analisi della approssimità

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

- *Nominale*. Una variabile può essere trattata come nominale quando i relativi valori rappresentano categorie prive di classificazione intrinseca (ad esempio il reparto della società in cui lavora un dipendente). Degli esempi di variabili nominali includono la regione, il codice postale e l'affiliazione religiosa.
- *Ordinale*. Una variabile può essere trattata come ordinale quando i suoi valori rappresentano categorie con qualche classificazione intrinseca (ad esempio i livelli di soddisfazione per un servizio, da molto insoddisfatto a molto soddisfatto). Degli esempi di variabili ordinali includono i punteggi di atteggiamento che rappresentano i gradi di soddisfazione o fiducia e i punteggi di classificazione delle preferenze.
- *Scala*. Una variabile può essere considerata di scala (continua) quando i relativi valori rappresentano categorie ordinate con una metrica significativa, tale che le comparazioni fra le distanze dei relativi valori siano appropriate. 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 della approssimità. La procedura presume che a ogni variabile sia stato assegnato il livello di misurazione appropriato; tuttavia, è possibile modificare temporaneamente il livello di misurazione per una variabile facendo clic con il tasto destro del mouse sulla variabile nell'elenco di variabili di origine e selezionando un livello di misurazione dal menu a comparsa.

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

Tabella 1. Icone del livello di misurazione

	Numerico	**Stringa**	**Data**	**Ora**
Scala (continuo)		n/d		

Tabella 1. Icone del livello di misurazione (Continua)

	Numerico	Stringa	Data	Ora
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.

Tutte le codifiche one-of-*c* si basano sui dati di training, anche se è definito un campione holdout (consultare "Partizioni" a pagina 90). Pertanto, se il campione holdout contiene casi con categorie di predittori assenti nei dati di training, non è possibile calcolarne il punteggio. Se il campione holdout 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. Tutta la modifica della scala viene eseguita in base ai dati di training, anche se è definito un campione holdout (consultare "Partizioni" a pagina 90). Se si specifica una variabile per definire le partizioni, è importante che tali funzioni abbiano distribuzioni simili nei campioni di training e holdout. Ad esempio, utilizzare la procedura Esplora per esaminare le distribuzioni nelle partizioni.

Peso della frequenza. I pesi della frequenza vengono ignorati da questa procedura.

Replica risultati. La procedura utilizza la generazione di numeri random durante l'assegnazione random delle partizioni e delle occorrenze con convalida incrociata. Se si desidera replicare in modo esatto i risultati, oltre ad utilizzare le impostazioni della stessa procedura, impostare un seme per Mersenne Twister (consultare "Partizioni" a pagina 90) oppure utilizzare le variabili per definire le partizioni e le occorrenze con convalida incrociata.

Per ottenere un'analisi della approssimità

Dai menu, scegliere:

Analizza > **Classifica** > **Vicino più vicino...**

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

   Destinazione (facoltativo). Se non è specificata alcuna destinazione (risposta o variabile dipendente), la procedura trova solo i *k* vicini più vicini, non viene eseguita alcuna classificazione o previsione.

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

Identificativo dei casi focali (facoltativo). Consente di contrassegnare casi di particolare interesse. Ad esempio, un ricercatore desidera determinare se i punteggi del test da un distretto scolastico, il caso focale, possono essere confrontati con quelli di distretti scolastici simili. Ricorrerà all'analisi della approssimità per individuare i distretti scolastici con le maggiori analogie in termini di uno specifico insieme di funzioni. Procederà quindi al confronto tra i punteggi del 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 di peer 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 (facoltativo). Ai casi vengono applicate etichette utilizzando questi valori nel grafico dello spazio delle funzioni, in quello dei peer 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) del dataset è sconosciuto. Poiché influisce sul calcolo dei risultati di questa procedura, il livello di misurazione deve essere definito per tutte le variabili.

Esegui scansione dati. Legge i dati del dataset attivo e assegna un livello di misurazione predefinito a tutti i campi con livello di misurazione sconosciuto. Con dataset 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 Vista variabile 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

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 la selezione delle funzioni non è attiva, viene utilizzata la convalida incrociata con occorrenze *V* per selezionare il “miglior” numero di vicini. Per informazioni sul controllo dell'assegnazione delle occorrenze, 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, delle differenze al quadrato 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 pesare 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 ripesando i valori di importanza della funzione in modo che la somma sia pari a 1.

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

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.

Criteri di arresto. A ogni fase, 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 di scala). La selezione in avanti prosegue fino al raggiungimento della condizione specificata.

- **Numero specificato di funzioni.** 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.
- **Modifica 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 diminuzione dei valori della modifica tenderà ad includere altre funzioni, con il rischio di includere funzioni che non aggiungono molto 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 modifica minima dipenderà dai dati e dall'applicazione. Per assistenza nella valutazione delle funzioni più importanti, vedere il log degli errori relativi alla selezione delle funzioni nell'output. Per ulteriori informazioni, consultare l'argomento “Log degli errori relativi alla selezione delle funzioni” a pagina 95.

Partizioni

La scheda Partizioni consente di suddividere il dataset in sottoinsiemi di training e holdout e, se possibile, di assegnare casi a occorrenze con convalida incrociata.

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

- **Assegna in modo random i casi alle partizioni.** Specificare la percentuale di casi da assegnare al campione di training. Il resto viene assegnato al campione holdout.

- **Usa variabile per assegnare casi.** Specificare una variabile numerica che assegni ogni caso nel dataset attivo al campione di training o holdout. I casi con valore positivo nella variabile vengono assegnati al campione di training, quelli con valore pari a 0 o negativo al campione holdout. I casi con un valore mancante di sistema vengono esclusi dall'analisi. I valori mancanti definiti dall'utente per la variabile partizione sono sempre considerati validi.

Occorrenze con convalida incrociata. La convalida incrociata con occorrenze *V* viene utilizzata per determinare il "miglior" numero di vicini. Per motivi legati alla prestazione, la convalida incrociata non è disponibile se si utilizza la selezione delle funzioni.

La convalida incrociata suddivide il campione in una serie di sottocampioni o occorrenze. 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 nella prima occorrenza del campione, il secondo modello si basa su tutti i casi eccetto quelli contenuti nella seconda occorrenza del campione 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 tutte le occorrenze.

- **Assegna in modo random i casi alle occorrenze.** Specificare il numero di occorrenze da utilizzare per la convalida incrociata. I casi vengono assegnati in modo casuale alle occorrenze, numerati da 1 a *V*, il numero di occorrenze.
- **Usa variabile per assegnare casi.** Specificare una variabile numerica che assegni ogni caso nel dataset attivo a un'occorrenze. La variabile deve essere un valore numerico compreso tra 1 e *V*. Se all'interno di tale intervallo mancano valori, e in corrispondenza delle suddivisioni in caso di file di suddivisione, si verificherà un errore.

Imposta seme per Mersenne Twister. Impostando un seme è possibile replicare le analisi. L'utilizzo di questo controllo è simile all'impostazione di Mersenne Twister come generatore attivo e alla specifica di un punto iniziale fisso nella finestra di dialogo Generatori di numeri random, con la differenza importante che impostando il seme in questa finestra di dialogo preserverà lo stato corrente del generatore di numeri random e ripristinerà tale stato una volta completata l'analisi.

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 di 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/holdout.** Se i casi vengono assegnati in modo casuale ai campioni training e holdout nella scheda Partizioni, viene salvato il valore della partizione (di training o holdout) a cui il caso è stato assegnato.
- **Variabile occorrenza con convalida incrociata.** Se nella scheda Partizioni alle occorrenze con convalida incrociata vengono assegnati casi in modo casuale, viene salvato il valore dell'occorrenza a cui è stato assegnato il caso.

Output

Output visualizzatore

- **Riepilogo elaborazione 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 holdout.
- **Grafici e tabelle.** Visualizza l'output relativo al modello, tra cui tabelle e grafici. Le tabelle nella vista modello comprendono *k* vicini più vicini e le distanze per i casi focali, la classificazione delle variabili di risposta categoriali e un riepilogo degli errori. L'output grafico nella vista modello include un log degli errori relativi alla selezione, il grafico dell'importanza delle funzioni, quello dello spazio delle funzioni e dei peer e la mappa dei quadranti. Per ulteriori informazioni, consultare l'argomento "Vista Modello".

File

- **Esporta modello in file XML.** È possibile utilizzare questo file 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 di suddivisione.
- **Esporta distanze tra casi focali e k vicini più vicini.** Per ogni caso focale viene creata una variabile distinta per ciascun *k* vicino più vicino (dal campione di training) del caso focale stesso e delle corrispondenti *k* distanze più vicine.

Opzioni

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 Modello

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

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

  La vista modello 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 pannello 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.

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

Spazio di funzioni

Il grafico dello spazio delle 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 l'ubicazione dei punti nel grafico indica i valori di tali funzioni per i casi nelle partizioni di training e holdout.

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

- La forma indica la partizione (Training o Holdout) 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 holdout è indicato quello previsto. Se non viene specificato alcun obiettivo, questo simbolo non viene visualizzato.
- Schemi più marcati indicano che un caso è focale. I casi focali vengono visualizzati collegati ai relativi *k* vicini più vicini.

Controlli 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" sono semplicemente i punti selezionati nel grafico 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 controlli "normali" per la selezione del punto; facendo clic su un punto, questo viene selezionato e vengono deselezionati tutti gli altri; facendo clic su un punto tenendo premuto il tasto Ctrl, lo si aggiunge all'insieme di punti selezionati. Le viste collegate, ad esempio il grafico dei peer, 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 visualizzato un suggerimento 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 "Reimposta" consente di ripristinare lo stato originale dello Spazio di funzioni.

Aggiunta e rimozione di campi/variabili

È possibile aggiungere nuovi campi/nuove variabili allo spazio di funzioni o rimuovere quelli attualmente visualizzati.

Palette Variabili

La palette Variabili deve essere visualizzata prima di poter aggiungere e rimuovere le variabili. Per visualizzare la palette Variabili, il Visualizzatore modelli deve essere in modalità Modifica e un caso deve essere selezionato nello spazio di funzioni.

1. Per impostare il Visualizzatore modelli in modalità Modifica, dal menu scegliere:
   Visualizza > **Modalità Modifica**
2. Una volta in Modalità Modifica, fare clic su un caso nello spazio di funzioni.
3. Per visualizzare la palette Variabili, dai menu scegliere:
   Visualizza > **Palette** > **Variabili**
   La palette Variabili elenca tutte le variabili nello spazio di funzioni. L'icona accanto al nome della variabile indica il livello di misurazione della variabile.
4. Per cambiare temporaneamente il livello di misurazione di una variabile, fare clic con il tasto destro del mouse sulla variabile nella palette Variabili e scegliere un'opzione.

Zone delle variabili

Le variabili vengono aggiunte alle "zone" nello spazio di funzioni. Per visualizzare le zone, iniziare a trascinare una variabile dalla palette Variabili oppure selezionare **Mostra zone**.

Lo spazio di funzioni ha delle zone per gli assi *x*, *y* e *z*.

Spostamento delle variabili nelle zone

Di seguito vengono fornite alcune regole generali e suggerimenti per lo spostamento delle variabili nelle zone.

- Per spostare una variabile in una zona, fare clic e trascinare la variabile dalla palette Variabili e rilasciarla nella zona. Se si sceglie **Mostra zone**, è anche possibile fare clic con il tasto destro del mouse su una zona e selezionare una variabile che si desidera aggiungere alla zona.
- Se si trascina una variabile dalla palette Variabili in una zona già occupata da un'altra variabile, la vecchia variabile viene sostituita con quella nuova
- Se si trascina una variabile da una zona in una zona già occupata da un'altra variabile, le variabili si scambiano posizione.
- Facendo clic su X in una zona, la variabile viene rimossa dalla zona in questione.
- Se vi sono più elementi grafici nella visualizzazione, ciascun elemento grafico può avere le proprie zone delle variabili associate. Selezionare prima l'elemento grafico.

Importanza della variabile

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

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

Collegamenti. Il grafico dei peer è collegato allo Spazio di funzioni in due modi.

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

Distanze dei vicini più vicini

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 identificativo 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 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

Il grafico mostra i casi focali e i relativi *k* vicini più vicini 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 pannelli 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 variabile vengono tracciate righe di riferimento.

Log degli errori relativi alla selezione delle 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.

Log degli errori relativi alla selezione di 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*.

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

Si tratta di grafici di selezione delle funzioni (consultare "Log degli errori relativi alla selezione delle funzioni"), a pannelli per *k*. Questo grafico è disponibile se esiste una destinazione e la selezione funzioni e la selezione *k* sono entrambe attive.

Tabella di classificazione

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 holdout contiene casi di holdout con valori mancanti sull'obiettivo. Questi casi contribuiscono al campione di Holdout: valori di Percentuale globale, ma non ai valori di Percentuale di correttezza.

Riepilogo degli errori

La tabella è disponibile in presenza di una variabile di destinazione. Visualizza l'errore associato al modello; la somma dei quadrati per una destinazione continua e il tasso di errore (100% – percentuale globale di correttezza) per una destinazione categoriale.

Capitolo 21. Analisi discriminante

L'analisi discriminante crea un modello di previsione per l'appartenenza al gruppo. 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 per cui è nota l'appartenenza al gruppo; è quindi possibile applicare le funzioni ai nuovi casi con misurazioni per le variabili predittore, ma di cui non è nota l'appartenenza al gruppo.

Nota: la variabile di raggruppamento può avere 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 di popolamento 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 a fasi, nella funzione non si dovranno includere tutte e quattro le variabili.

Statistiche. Per ogni variabile: medie, deviazioni standard, ANOVA univariate. Per ogni analisi: *M* di Box, matrice di correlazione entro gruppi, matrice di covarianza entro gruppi, matrice di covarianza di gruppi separati, matrice di covarianza totale. Per ogni funzione discriminante canonica: autovalore, percentuale di varianza, correlazione canonica, lambda di Wilks, chi-quadrato. Per ogni fase: probabilità a priori, coefficienti di funzione di Fisher, coefficienti di funzione non standardizzati, lambda di Wilks per ogni funzione canonica.

Considerazioni sui dati relative all'analisi discriminante

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 dummy o di contrasto.

Ipotesi. I casi devono essere indipendenti. Le variabili predittore 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

1. Dai menu, scegliere:
   Analizza > **Classifica** > **Discriminante...**

2. Selezionare una variabile di raggruppamento con valori interi e fare clic su **Definisci intervallo** per specificare le categorie desiderate.
3. 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).
4. 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 a fasi.** Usa l'analisi a fasi per controllare l'immissione della variabile e la rimozione.
5. È inoltre possibile selezionare i casi utilizzando una variabile di selezione.

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

Per selezionare i casi da usare nell'analisi:

1. Scegliere la variabile di selezione nella finestra di dialogo Analisi discriminante.
2. Fare clic su**Valore** per immettere un intero come variabile di selezione.

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 addestramento e test con i quali eseguire la convalida del modello generato.

Analisi discriminante: Statistiche

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

- *Medie*. Visualizza le medie globali e di gruppo e le e deviazioni standard per le variabili indipendenti.
- *ANOVA univariate*. Esegue un test di analisi della varianza a una via dell'uguaglianza delle medie di gruppo per ogni variabile indipendente.
- *M di Box*. Un test per l'uguaglianza di matrici di covarianza di gruppo. Per dimensioni del 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.

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

- *di Fisher*. Visualizza i coefficienti di Fisher della funzione discriminante, che possono essere usati direttamente per la classificazione. Viene ottenuto un insieme separato di coefficienti di funzioni di classificazione per ciascun gruppo e a 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 una matrice di correlazione entro gruppi raggruppata ottenuta calcolando la media delle matrici di covarianza separate per tutti i gruppi prima di calcolare le correlazioni.
- *Covarianza entro gruppi*. Visualizza una matrice di covarianza entro gruppi raggruppata, che può essere differente dalla matrice di covarianza totale. La matrice viene ottenuta calcolando la media delle matrici di covarianza separate per tutti i gruppi.
- *Covarianza di gruppi separati*. Visualizza le matrici di covarianza separate per ciascun gruppo.
- *Covarianza totale*. Visualizza una matrice di covarianza da tutti i casi come se fossero da un singolo campione.

Analisi discriminante: Metodo a fasi

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 a fasi che sceglie le variabili da inserire nell'equazione in base a quanto esse contribuiscono a minimizzare il Lambda di Wilks. A ogni fase, viene inserita la variabile che minimizza il Lambda di Wilks globale.
- *Varianza non spiegata*. Ad ogni fase viene inserita la variabile che riduce al minimo la somma della variazione spiegata fra gruppi.
- *Distanza di Mahalanobis*. Una misura di quanto i valori di un caso sulle variabili indipendenti differiscono dalla media di tutti i casi. Un'elevata distanza di Mahalanobis identifica un caso come caratterizzato da valori estremi per una o più variabili indipendenti.
- *Rapporto F più piccolo*. Un metodo di selezione delle variabili nelle analisi a fasi basato sulla massimizzazione di un rapporto F calcolato dalla distanza di Mahalanobis tra gruppi.
- *V di Rao*. Una misura delle differenze tra medie di gruppo. Detta anche traccia di Lawley-Hotelling. Ad ogni fase viene inserita la variabile che massimizza l'aumento della V di Rao. Dopo aver selezionato questa opzione, immettere il valore minimo che una variabile deve avere per essere inserita nell'analisi.

Criteri. Le alternative disponibili sono **Usa valore di F** e **Usa probabilità di F**. Specificare i valori per immettere e rimuovere le 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 Per rimuovere un numero maggiore di variabili dal modello, aumentare il valore di rimozione.
- *Usa probabilità di F*. Una 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 Per rimuovere un numero maggiore di variabili dal modello, ridurre il valore di rimozione.

Visualizza. Riepilogo delle fasi visualizza le statistiche per tutte le variabili dopo ogni fase; **F per distanze a coppie** visualizza una matrice di rapporti *F* a coppie per ogni coppia di gruppi.

Analisi discriminante: Classificazione

Probabilità a priori. Questa opzione determina se i coefficienti di classificazione vengono adattati per una conoscenza a priori dell'appartenenza al gruppo.

- **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 dell'appartenenza al gruppo. 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 adattati in modo da aumentare la probabilità di appartenenza nel primo gruppo rispetto agli altri due.

Visualizza. Le opzioni di visualizzazione disponibili sono: risultati per casi, tabella riassuntiva e classificazione leav-one-out.

- *Risultati per casi*. Per ciascun caso vengono visualizzati il gruppo effettivo, il gruppo previsto, le probabilità posteriori e i punteggi discriminanti.
- *Tabella Riepilogo*. 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 leave-one-out*. Ogni caso nell'analisi viene classificato usando le funzioni derivate da tutti i casi meno se stesso. È noto anche come "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 gruppi.
- *Gruppi separati*. Le matrici di covarianza dei gruppi separati sono usate per la classificazione. Poiché la classificazione è basata sulle funzioni discriminanti (non sulle variabili 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 combinati*. Crea un grafico a dispersione di tutti i gruppi dei primi due valori di funzione discriminante. Se c'è una sola funzione, viene invece visualizzato un istogramma.
- *Gruppi separati*. Crea dei grafici a dispersione dei gruppi separati dei primi due valori di funzione discriminante. Se c'è una sola funzione, vengono invece visualizzati degli istogrammi.
- *Mappa territoriale*. Un grafico dei limiti 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 limiti. La mappa non viene visualizzata se c'è una sola funzione discriminante.

Analisi discriminante: Salva

È possibile aggiungere nuove variabili al file di dati attivo. Le opzioni disponibili sono: appartenenza al gruppo 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 anche possibile esportare le informazioni sul modello nel file specificato in formato XML. È possibile utilizzare questo file 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 fattoriali, che possono essere salvati come variabili per le analisi successive.

Esempio. Quali sono gli atteggiamenti sottostanti che inducono le persone a rispondere a un'indagine politica in un determinato modo? Dall'esame delle correlazioni esistenti tra le voci dell'indagine 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 ogni analisi del fattore: matrice di correlazione delle variabili, tra cui i livelli di significatività, determinante e inverso: matrice di correlazione riprodotta, tra cui anti-immagine; soluzione iniziale (comunalità, autovalori e percentuale di varianza spiegata); misura di Kaiser-Meyer-Olkin di adeguatezza del campionamento e test di sfericità di Bartlett; soluzione non ruotata, tra cui caricamenti dei fattori, comunalità e autovalori; soluzione ruotata, tra cui matrice del modello ruotato e matrice di trasformazione. Per le rotazioni oblique: modello ruotato e matrici della struttura; matrice del coefficiente di punteggio del fattore e matrice di covarianza del fattore. Grafici: grafico scree delle eguaglianze e grafico di caricamento dei primi due o tre fattori.

Considerazioni sui dati relative all'analisi fattoriale

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.

Ipotesi. I dati devono avere una distribuzione normale bivariata per ogni 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

1. Dai menu, scegliere:
   Analizza > **Riduzione delle dimensioni** > **Fattore...**

2. Selezionare le variabili per l'analisi fattoriale.

Analisi fattoriale: Seleziona casi

Per selezionare i casi da usare nell'analisi:

1. Selezionare una variabile di selezione.
2. Fare clic su**Valore** per immettere un intero come variabile di selezione.

Nell'analisi fattoriale verranno utilizzati solo i casi con tale valore per la variabile di selezione.

Analisi fattoriale: Descrittive

Statistiche. Descrittive univariate include la media, la deviazione standard e il numero di casi validi per ogni 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 di campionamento 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 che il modello fattoriale è inappropriato.
- *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 negativi dei coefficienti di correlazione parziale e la matrice di covarianza anti-immagine contiene i negativi delle covarianze parziali. In un buon modello fattoriale, la maggior parte degli elementi fuori dalla diagonale dovrebbe avere valori bassi. La misura dell'adeguatezza del campionamento per una variabile è visualizzata sulla diagonale della matrice di correlazione anti-immagine.

Analisi fattoriale: Estrazione

Metodo. Consente di specificare il metodo di estrazione del fattore. I metodi disponibili sono: componenti principali, minimi quadrati non pesati, minimi quadrati generalizzati, massima verosimiglianza, fattorizzazione dell'asse principale, fattorizzazione alfa e fattorizzazione immagine.

- *Analisi componenti principali*. Un metodo di estrazione dei fattori 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 minimi quadrati non pesati*. Un metodo di estrazione dei fattori che minimizza la somma delle differenze al quadrato tra la matrice di correlazione osservata e quella riprodotta (ignorando le diagonali).
- *Metodo minimi quadrati generalizzato*. Un metodo di estrazione dei fattori che minimizza la somma delle differenze al quadrato tra la matrice di correlazione osservata e la matrice di correlazione riprodotta. Le correlazioni sono pesate tramite l'inverso della loro unicità, in modo da dare meno peso alle variabili con elevata unicità rispetto a quelle con unicità inferiore.
- *Metodo di massima verosimiglianza*. Un metodo di estrazione dei fattori che produce le stime del parametro che più verosimilmente hanno prodotto la matrice di correlazione osservata, se il campione è estratto da una distribuzione normale multivariata. Le correlazioni sono pesate in base all'inverso dell'unicità delle variabili; viene usato 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 nella diagonale come stime iniziali

delle comunalità. Questi caricamenti fattori 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 dall'universo delle variabili potenziali. Questo metodo massimizza l'affidabilità alfa dei fattori.
- *Fattorizzazione immagine*. Un 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 una 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.

Visualizza. Consente di richiedere la soluzione fattoriale non ruotata e un grafico scree.

- *Soluzione fattoriale non ruotata*. Visualizza i caricamenti fattori non ruotati (matrice del modello fattoriale), le comunalità e gli autovalori per la soluzione fattoriale.
- *Grafico scree*. Un grafico della varianza associata a ciascun fattore. Questo grafico viene usato per determinare il numero di fattori da mantenere. Di norma, il grafico mostra una distinta interruzione tra la brusca pendenza dei fattori di grandi dimensioni e la graduale sequenza del resto (lo scree).

Numero massimo di iterazioni per la convergenza. Consente di specificare il numero massimo di fasi che l'algoritmo può eseguire per valutare la soluzione.

Analisi fattoriale: Rotazione

Metodo. Consente di selezionare il metodo di rotazione fattoriale. I metodi disponibili sono: varimax, equamax, quartimax, oblimin diretto e promax.

- *Metodo Varimax*. Un metodo di rotazione ortogonale che minimizza il numero di variabili che hanno elevati caricamenti su ciascun fattore. Questo metodo semplifica l'interpretazione dei fattori.
- *Metodo Oblimin diretto*. Un metodo per la rotazione obliqua (non ortogonale). Quando delta vale 0 (il valore predefinito), le soluzioni sono per la maggior parte oblique. Quando delta diventa negativo e aumenta in valore assoluto, i fattori cominciano a essere meno obliqui. Per sovrascrivere il delta predefinito di 0, immettere un numero inferiore o uguale a 0,8.
- *Metodo Quartimax*. Un metodo di rotazione che minimizza il numero di fattori necessari per spiegare ogni variabile. Questo metodo semplifica l'interpretazione delle variabili osservate.
- *Metodo equamax*. Un metodo di rotazione che è una combinazione del metodo varimax, che semplifica i fattori, e del metodo quartimax, che semplifica le variabili. Il numero di variabili con un notevole carico su un fattore e il numero di fattori necessario per spiegare una variabile sono minimizzati.
- *Rotazione Promax*. Una rotazione oblique che consente di correlare i fattori. Questa rotazione può essere calcolata più rapidamente di una rotazione Oblimin diretta ed è pertanto utile per dataset di grandi dimensioni.

Visualizza. Consente di includere l'output nella soluzione ruotata, nonché i grafici dei caricamenti per i primi due o tre fattori.

- *Soluzione ruotata*. Per ottenere una soluzione ruotata deve essere 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 caricamenti fattori*. Grafico tridimensionale del caricamento 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.

Numero massimo di iterazioni per la convergenza. Consente di specificare il massimo numero di fasi che l'algoritmo può eseguire per completare la rotazione.

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 di stima dei coefficienti di punteggio fattoriale. I punteggi prodotti hanno media 0 e varianza pari al quadrato della correlazione multipla fra i punteggi fattoriali stimati e i valori reali dei fattori. I punteggi possono essere correlati anche quando i fattori sono ortogonali.
- *Punteggi Bartlett*. Un metodo di stima dei coefficienti di punteggio fattoriale. I punteggi prodotti hanno una media pari a 0. La somma dei quadrati dei fattori univoci sull'intervallo delle variabili è minimizzata.
- *Metodo di Anderson-Rubin*. Un metodo per stimare i coefficienti dei punteggi fattoriali; una modifica del metodo di Bartlett che garantisce l'ortogonalità dei fattori stimati. I punteggi prodotti hanno una media pari a 0, una deviazione standard pari a 1 e non sono correlati.

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

Valori mancati. Consente di specificare le modalità di gestione dei valori mancanti. Le scelte disponibili sono: Escludi casi *a livello di elenco*, Escludi casi *a coppie* 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.

Funzioni 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 fattoriali 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.

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

Capitolo 23. Scelta di una procedura per il raggruppamento

È possibile eseguire cluster analysis utilizzando le procedure Analisi Cluster TwoStep, Cluster gerarchica o Cluster 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.

Analisi cluster gerarchica. La procedura Analisi 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 Analisi cluster gerarchica analizza variabili per intervallo (continue), binarie o di conteggio, purché le variabili siano tutte dello stesso tipo.

Analisi del cluster delle k Medie. La procedura Analisi del cluster 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 eseguire la lettura dai centri cluster iniziali e salvare i centri cluster finali in un file esterno 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 dataset, che non sarebbero altrimenti evidenti. L'algoritmo utilizzato da questa procedura presenta diverse funzioni 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 una struttura ad albero delle funzioni cluster (o CF/cluster feature) 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.

Misura di distanza. Questa selezione determina la modalità di calcolo della similarità tra due cluster.

- **Verosimiglianza logaritmica.** 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 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 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, consultare l'argomento “Opzioni di Analisi cluster TwoStep” a pagina 110.

Criterio di clustering. Questa selezione determina la modalità di definizione del numero dei cluster mediante l'algoritmo di cluster 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).

Considerazioni sui dati dell'analisi cluster TwoStep

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 struttura ad albero delle funzioni 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 random 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 random diversi.

Ipotesi. La misura della 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 ipotesi di indipendenza sia delle ipotesi di distribuzione, ma è consigliabile verificare fino a che punto tali ipotesi vengano soddisfatte.

Utilizzare la procedura Correlazioni bivariate per testare l'indipendenza di due variabili continue. Utilizzare la procedura Tabelle di contingenza per verificare l'indipendenza di due variabili categoriali. Utilizzare la procedura Medie per testare l'indipendenza tra una variabile continua e la variabile categoriale. Utilizzare la procedura Esplora per testare la normalità di una variabile continua. Utilizzare la procedura Test del chi-quadrato per testare se una variabile categoriale ha una distribuzione multinomiale specificata.

Per ottenere un'analisi cluster TwoStep

1. Dai menu, scegliere:

   Analizza > **Classifica** > **Cluster TwoStep...**
2. Selezionare una o più variabili categoriali o continue.

Se lo si desidera, è possibile:

- Adattare 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 del Visualizzatore modelli.
- Salvare i risultati del modello nel file di lavoro o in un file XML esterno.

Opzioni di Analisi cluster TwoStep

Trattamento valori anomali. Questo gruppo consente di considerare i valori anomali in particolare durante il clustering, se la struttura ad albero delle funzioni cluster (CF) è piena. La struttura ad albero delle funzioni cluster (CF) è piena 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 la struttura ad albero delle funzioni cluster (CF) risulta piena, 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 ad albero, i valori anomali vengono inseriti nella struttura ad albero delle funzioni cluster (CF), se possibile. Altrimenti, vengono eliminati.
- Se non si seleziona la gestione del rumore e la struttura ad albero delle funzioni cluster (CF)risulta piena, 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 di valori anomali viene assegnato un numero di identificazione –1 e non viene incluso nel conteggio del numero di 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 di sistema.
- L'algoritmo potrebbe non riuscire a trovare il numero di cluster corretto o specificato se questo valore è troppo basso.

Standardizzazione della variabile. L'algoritmo di cluster 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 della struttura ad albero delle funzioni cluster (CF). Le seguenti impostazioni dell'algoritmo di cluster riguardano in modo specifico la struttura ad albero delle funzioni cluster (CF) e devono essere modificate con cautela:

- **Soglia modifica distanza iniziale.** Si tratta della soglia iniziale utilizzata per ampliare la struttura ad albero delle funzioni cluster (CF). Se dopo l'inserimento di un determinato caso in una foglia della struttura ad albero delle funzioni cluster (CF), 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 ad albero.** Il numero massimo di livelli della struttura ad albero delle funzioni cluster (CF).
- **Massimo numero di nodi possibile.** Indica il numero massimo di nodi della struttura ad albero delle funzioni cluster (CF) che questa procedura è in grado di generare, in base alla funzione ($b^{d+1} - 1) / (b - 1$), dove b rappresenta le ramificazioni massime e d è la profondità massima della struttura ad albero. Tenere presente che una struttura ad albero delle funzioni cluster (CF), di dimensioni eccessive può rappresentare un peso considerevole per le risorse del sistema e compromettere quindi la prestazione della procedura. Ogni nodo richiede un minimo di 16 byte.

Aggiornamento del modello di cluster. Questo gruppo consente di importare e aggiornare un modello di cluster generato da un'analisi precedente. Il file di input contiene la struttura ad albero delle funzioni cluster (CF) 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, consultare l'argomento “Output di Analisi cluster TwoStep” a pagina 112.

Se viene selezionato l'aggiornamento del modello di cluster, verranno utilizzate le opzioni per la generazione della struttura ad albero delle funzioni cluster (CF) specificate per il modello originale. Vengono quindi utilizzate le impostazioni del modello salvato relative a misura della distanza, gestione del rumore, allocazione di memoria e ottimizzazione della struttura ad albero delle funzioni cluster (CF), mentre qualsiasi nuova impostazione specificata nelle finestre di dialogo viene ignorata.

Nota: quando si esegue un aggiornamento del modello di cluster, la procedura presume che nessuno dei casi selezionati nel dataset 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

Output. Questo gruppo fornisce le opzioni per la visualizzazione dei risultati dei raggruppamenti.

- **Tabelle pivot.** I risultati vengono visualizzati nelle tabelle pivot.
- **Grafici e tabelle in Visualizzatore modelli.** I risultati vengono visualizzati nel Visualizzatore modelli.
- **Campi valutazione.** Calcola i dati dei cluster per le variabili che non sono stati utilizzati nella creazione dei cluster. I campo valutazione possono essere visualizzati insieme alle funzioni di input nel Visualizzatore modelli selezionandoli nella finestra di dialogo secondaria Visualizza. I campi con valori mancanti vengono ignorati.

File di dati di lavoro. Questo gruppo consente di salvare le variabili all'interno del dataset attivo.

- **Crea variabile di appartenenza cluster.** Questa variabile contiene un numero di identificazione del cluster per ogni caso. Il nome di questa variabile è *tsc_n*, dove *n* è un intero positivo che indica l'ordinale dell'operazione di salvataggio del dataset attivo eseguita mediante questa procedura in una determinata sessione.

File XML. Il modello di cluster finale e la struttura ad albero delle funzioni cluster (CF) 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 modello per applicare le informazioni del modello ad altri file di dati per il calcolo del punteggio.
- **Esporta struttura ad albero delle funzioni cluster (CF).** Questa opzione consente di salvare lo stato corrente della struttura ad albero del cluster e aggiornarlo in un secondo momento utilizzando dati più aggiornati.

Il Visualizzatore cluster

I modelli di cluster in genere vengono utilizzati per trovare gruppi (o cluster) di record simili in base alle variabili esaminate, dove la somiglianza tra i membri dello stesso gruppo è elevata e la somiglianza tra i membri di gruppi differenti è bassa. I risultati possono essere utilizzati per identificare le associazioni che altrimenti non sarebbero visibili. Ad esempio, mediante l'analisi cluster delle preferenze dei clienti, del livello di reddito e delle abitudini di acquisto è possibile identificare i tipi di clienti che risponderanno con maggiore probabilità a una determinata campagna di marketing.

Sono disponibili due approcci per interpretare i risultati in una visualizzazione cluster.

- Esaminare i cluster per determinare le caratteristiche univoche del cluster in questione. *Un cluster contiene tutti coloro che chiedono denaro in prestito con reddito elevato? Questo cluster contiene più record di altri?*
- Esaminare i campi nei cluster per determinare come sono distribuiti i valori nei cluster. *Il livello di istruzione determina l'appartenenza a un cluster? Un punteggio di credito elevato distingue tra l'appartenenza a un cluster o a un altro?*

Utilizzando le viste principali e le varie viste collegate nel Visualizzatore cluster, è possibile ottenere informazioni utili per rispondere a queste domande.

Per informazioni sul modello di cluster, attivare (fare doppio clic) l'oggetto Visualizzatore modelli nel Visualizzatore.

Visualizzatore cluster

Il Visualizzatore cluster è composto da due pannelli, la vista principale sul lato sinistro e la vista collegata, o ausiliaria, sul lato destro. Vi sono due viste principali:

- Riepilogo del modello (valore predefinito). Per ulteriori informazioni, consultare l'argomento "Vista Riepilogo del modello" a pagina 113.

- Raggruppamenti. Per ulteriori informazioni, consultare l'argomento “Vista Cluster”.

Le viste collegate/ausiliarie sono quattro:

- Importanza predittore. Per ulteriori informazioni, consultare l'argomento “Visualizzazione Importanza predittore nei cluster” a pagina 115.
- Dimensioni cluster (valore predefinito). Per ulteriori informazioni, consultare l'argomento “Vista Dimensioni cluster” a pagina 115.
- Distribuzione cella. Per ulteriori informazioni, consultare l'argomento “Vista Distribuzione delle celle” a pagina 115.
- Comparazione tra cluster. Per ulteriori informazioni, consultare l'argomento “Vista Comparazione tra cluster” a pagina 115.

Vista Riepilogo del modello

La vista Riepilogo del modello mostra un'istantanea, o riepilogo, del modello di cluster, inclusa una misura Silhouette della coesione e separazione del cluster, che è ombreggiata per indicare risultati insufficienti, sufficienti o buoni. Questa istantanea consente di controllare rapidamente se la qualità è insufficiente, nel qual caso è possibile decidere di ritornare al nodo di modellazione per modificare le impostazioni del modello di cluster per ottenere un risultato migliore.

I risultati insufficienti, sufficienti e buoni si basano sul lavoro di Kaufman e Rousseeuw (1990) in merito all'interpretazione delle strutture di cluster. Nella vista Riepilogo del modello un buon risultato equivale ai dati che rappresentano la valutazione di Kaufman e Rousseeuw come prova ragionevole o forte della struttura del cluster, un risultato sufficiente rappresenta la loro valutazione della prova debole e un risultato insufficiente rappresenta la loro valutazione della prova non significativa.

La misura Silhouette calcola la media in tutti i record, (B−A) / max(A,B), in cui A è la distanza del record dal centro del suo cluster e B è la distanza del record dal centro del cluster più vicino a cui non appartiene. Il coefficiente silhouette pari a 1 indica che tutti i casi si trovano direttamente nei centri dei loro cluster. Il valore −1 indica che tutti i casi si trovano nei centri di altri cluster. Il valore 0 indica, in media, che i casi sono equidistanti tra il centro del proprio cluster e l'altro cluster più vicino.

Il riepilogo include una tabella contenente le informazioni riportate di seguito.

- **Algoritmo.** L'algoritmo di clustering utilizzato, ad esempio, "TwoStep".
- **Funzioni di input.** Il numero di campi, noti anche come **input** o **predittori**.
- **Cluster.** Il numero di cluster nella soluzione.

Vista Cluster

La vista Cluster contiene una griglia di cluster per funzioni che include i nomi, le dimensioni e i profili di ciascun cluster.

Le colonne della griglia contengono le informazioni riportate di seguito.

- **Cluster.** I numeri di cluster creati dall'algoritmo.
- **Etichetta.** Le etichette applicate a ciascun cluster (per impostazione predefinita, l'etichetta è vuota). Fare doppio clic nella cella per immettere un'etichetta che descriva il contenuto del cluster; ad esempio, "Acquirenti di automobili di lusso".
- **Descrizione.** Una descrizione del contenuto del cluster (per impostazione predefinita, la descrizione è vuota). Fare doppio clic nella cella per immettere una descrizione del cluster; ad esempio "Professionisti cinquantacinquenni con un reddito superiore a $100.000".
- **Dimensione.** La dimensione di ogni cluster come percentuale del campione di cluster globale. Ogni cella di dimensione all'interno della griglia visualizza una barra verticale che mostra la percentuale della dimensione all'interno del cluster, una percentuale della dimensione in formato numerico e i conteggi dei casi del cluster.

- **Funzioni.** I predittori e gli input singoli, ordinati, per impostazione predefinita, in base all'importanza globale. Se delle colonne hanno la stessa dimensione, vengono visualizzate in ordine crescente dei numeri di cluster.

  L'importanza globale delle funzioni viene indicata dal colore dell'ombreggiatura dello sfondo della cella; la funzione più importante è la più scura; la funzione meno importante non è ombreggiata. Una guida sopra la tabella indica l'importanza collegata a ogni colore della cella della funzione.

Quando si passa il mouse su una cella, vengono visualizzati il nome completo/l'etichetta della funzione e il valore relativo all'importanza della cella. È possibile che vengano visualizzate altre informazioni, a seconda del tipo di vista e di funzione. Nella vista Centri cluster sono incluse le statistiche della cella e il valore della cella; ad esempio: "Media: 4.32". Per le funzioni categoriali la cella mostra il nome della categoria più frequente (modale) e la relativa percentuale.

Nella vista Cluster è possibile selezionare diversi modi per visualizzare le informazioni del cluster:

- Trasporre cluster e funzioni. Per ulteriori informazioni, consultare l'argomento "Trasponi cluster e funzioni".
- Ordinare le funzioni. Per ulteriori informazioni, consultare l'argomento "Ordina funzioni".
- Ordinare i cluster. Per ulteriori informazioni, consultare l'argomento "Ordina cluster".
- Selezionare il contenuto della cella. Per ulteriori informazioni, consultare l'argomento "Contenuto della cella".

Trasponi cluster e funzioni: Per impostazione predefinita, i cluster vengono visualizzati come colonne e le funzioni vengono visualizzate come righe. Per invertire questa visualizzazione, fare clic sul pulsante **Trasponi cluster e funzioni** sul lato destro dei pulsanti **Ordina funzioni per**. Ad esempio, questa operazione è utile se sono visualizzati molti cluster e si desidera ridurre lo scorrimento orizzontale necessario per visualizzare i dati.

Ordina funzioni: I pulsanti **Ordina funzioni per** consentono di selezionare come vengono visualizzate le celle delle funzioni.

- **Importanza globale.** Si tratta dell'ordinamento predefinito. Le funzioni vengono disposte in ordine decrescente di importanza globale e l'ordinamento è lo stesso nei cluster. Se delle funzioni hanno valori di importanza correlati, le funzioni correlate vengono elencate in ordine crescente dei nomi delle funzioni.
- **Importanza entro i cluster.** Le funzioni vengono ordinate rispetto alla loro importanza per ogni cluster. Se delle funzioni hanno valori di importanza correlati, le funzioni correlate vengono elencate in ordine crescente dei nomi delle funzioni. Quando si sceglie questa opzione, l'ordinamento in genere varia tra i cluster.
- **Nome.** Le funzioni vengono disposte per nome in ordine alfabetico.
- **Ordine dei dati.** Le funzioni vengono disposte in base al loro ordine nel dataset.

Ordina cluster: Per impostazione predefinita, i cluster sono disposti in ordine decrescente di dimensione. I pulsanti **Ordina cluster per** consentono di disporli per nome in ordine alfabetico o, se sono state create delle etichette univoche, in ordine alfanumerico di etichetta.

Le funzioni che hanno la stessa etichetta sono ordinate per nome cluster. Se i cluster sono ordinati per etichetta e si modifica l'etichetta di un cluster, l'ordinamento viene aggiornato automaticamente.

Contenuto della cella: I pulsanti **Celle** consentono di modificare la visualizzazione del contenuto della cella per i campi delle funzioni e di valutazione.

- **Centri cluster.** Per impostazione predefinita, le celle visualizzano i nomi/le etichette delle funzioni e la tendenza centrale per ogni combinazione di cluster/funzione. Per i campi continui viene visualizzata la media e per i campi categoriali viene mostrata la modalità (la categoria riscontrata più di frequente) con la percentuale della categoria.

- **Distribuzioni assolute.** Mostra le distribuzioni assolute dei nomi/delle etichette delle funzioni all'interno di ogni cluster. Per le funzioni categoriali, i grafici a barre vengono mostrati sovrapposti e le categorie sono disposte in ordine crescente dei valori dei dati. Per le funzioni continue, viene mostrato un grafico di densità livellato, che utilizza gli stessi endpoint e intervalli per ogni cluster.

  Il pannello rosso scuro mostra la distribuzione del cluster, mentre il pannello più chiaro rappresenta i dati globali.
- **Distribuzioni relative.** Mostra i nomi/le etichette della funzione e le distribuzioni relative nelle celle. In generale, le visualizzazioni sono simili a quelle mostrate per le distribuzioni assolute, con la differenza che vengono visualizzate le distribuzioni relative.

  Il pannello rosso scuro mostra la distribuzione del cluster, mentre il pannello più chiaro rappresenta i dati globali.
- **Vista di base.** Quando sono presenti molti cluster, può essere difficile vedere tutti i dettagli senza scorrere i dati. Per ridurre la necessità dello scorrimento, selezionare questa vista per modificare la visualizzazione su una versione più compatta della tabella.

Visualizzazione Importanza predittore nei cluster

La visualizzazione Importanza predittore mostra l'importanza relativa di ciascun campo nella stima del modello.

Vista Dimensioni cluster

La vista Dimensioni cluster mostra un grafico a torta contenente ogni cluster. La dimensione della percentuale di ogni cluster viene visualizzata in ogni sezione; passare il mouse su ogni sezione per visualizzare il conteggio nella sezione in questione.

Una tabella sotto il grafico elenca le informazioni sulla dimensione riportate di seguito.

- La dimensione del cluster più piccolo (il conteggio e la percentuale dell'insieme).
- La dimensione del cluster più grande (il conteggio e la percentuale dell'insieme).
- Il rapporto della dimensione del cluster più grande con il cluster più piccolo.

Vista Distribuzione delle celle

La vista Distribuzione delle celle mostra un grafico espanso e più dettagliato della distribuzione dei dati per le celle della funzione selezionate nella tabella nel pannello principale Cluster.

Vista Comparazione tra cluster

La vista Comparazione tra cluster è composta da un layout stile griglia con le funzioni nelle righe e i cluster selezionati nelle colonne. Questa vista consente di comprendere meglio i fattori che costituiscono i cluster; inoltre, consente di vedere le differenze tra i cluster non solo rispetto ai dati globali, ma tra di loro.

Per selezionare i cluster da visualizzare, fare clic nella parte superiore della colonna del cluster nel pannello principale Cluster. Utilizzare i tasti Ctrl-clic o Maiusc-clic per selezionare o deselezionare più di un cluster per la comparazione.

Nota: è possibile selezionare fino a 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 per**. Quando si seleziona **Importanza entro i cluster**, i campi vengono sempre ordinati per importanza globale.

I grafici di sfondo mostrano le distribuzioni globali di ciascuna funzione.

- Le funzioni categoriali vengono mostrate come grafici a punti, in cui la dimensione del punto indica la categoria più frequente/modale per ciascun cluster (per funzione).
- Le funzioni continue vengono visualizzate come grafici a scatole, che mostrano le mediane globali e gli intervalli interquartili.

Sovrapposti a queste viste di sfondo sono i grafici a scatole per i cluster selezionati:

- Per le funzioni continue, i contrassegni a punta quadrata e le linee orizzontali indicano la mediana e l'intervallo interquartile per ciascun cluster.
- Ogni cluster è rappresentato da un colore differente, mostrato nella parte superiore della vista.

Navigazione nel visualizzatore cluster

Il visualizzatore cluster è una visualizzazione interattiva. È possibile:

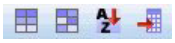
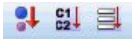
- Selezionare un campo o un cluster per visualizzare più dettagli.
- Confrontare i cluster per selezionare gli elementi a cui si è interessati.
- Modificare la visualizzazione.
- Trasporre gli assi.

Utilizzo delle barre degli strumenti

L'utente controlla le informazioni visualizzate nei pannelli sulla destra e sulla sinistra utilizzando le opzioni della barra degli strumenti. È possibile modificare l'orientamento della visualizzazione (dall'alto verso il basso, da sinistra a destra o da destra a sinistra) utilizzando i controlli della barra degli strumenti. Inoltre, è possibile ripristinare le impostazioni predefinite del visualizzatore e aprire una finestra di dialogo per specificare il contenuto della vista Cluster nel pannello principale.

Le opzioni **Ordina funzioni per**, **Ordina cluster per**, **Celle** e **Visualizzazione** sono disponibili solo quando si seleziona la vista **Cluster** nel pannello principale. Per ulteriori informazioni, consultare l'argomento “Vista Cluster” a pagina 113.

Tabella 2. Icone della barra degli strumenti.

Icona	Argomento
	Consultare Trasponi cluster e funzioni
	Consultare Ordina funzioni per
	Consultare Ordina cluster per
	Consultare Celle

Controllo della visualizzazione della vista cluster

Per controllare cosa viene visualizzato nella vista Cluster nel pannello principale, fare clic sul pulsante **Visualizza**; viene aperta la finestra Visualizza.

Funzioni. Opzione selezionata per impostazione predefinita. Per nascondere tutte le funzioni di input, deselezionare la casella di controllo.

Campi valutazione. Scegliere i campi di valutazione (campi non utilizzati per creare il modello cluster, ma inviati al visualizzatore modelli per valutare i cluster) da visualizzare; nessuno di essi viene visualizzato per impostazione predefinita. *Nota*: questa casella di controllo non è disponibile se non è disponibile alcun campo di valutazione.

Descrizioni cluster. Opzione selezionata per impostazione predefinita. Per nascondere tutte le celle di descrizione, deselezionare la casella di controllo.

Dimensioni cluster. Opzione selezionata per impostazione predefinita. Per nascondere tutte le celle di dimensione del cluster, deselezionare la casella di controllo.

Numero massimo di categorie. Specificare il numero massimo di categorie da visualizzare nei grafici delle funzioni categoriali; il valore predefinito è 20.

Filtraggio dei record

Se si desidera avere ulteriori informazioni sui casi in un determinato cluster o gruppo di cluster, è possibile selezionare un sottoinsieme di record per un'analisi più approfondita basata sui cluster selezionati.

1. Selezionare i cluster nella vista Cluster del Visualizzatore cluster. Per selezionare più cluster, utilizzare Ctrl-clic.
2. Dai menu, scegliere:

   Genera > **Filtra record...**
3. Inserire un nome per la variabile di filtro. I record dai cluster selezionati riceveranno un valore 1 per questo campo. Tutti gli altri record riceveranno il valore 0 e verranno esclusi dalle analisi successive finché non si modifica lo stato del filtro.
4. Fare clic su **OK**.

Capitolo 25. Analisi cluster gerarchica

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 Prossimità. 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 l'analisi 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. Pianificazione 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 ghiacciolo.

Considerazioni sui dati per l'Analisi cluster gerarchica

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 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 Analisi cluster gerarchica.

Ordine dei casi. Se le distanze assegnate o le similarità sono presenti nei dati di input o nei cluster aggiornati durante l'unione, la soluzione cluster risultante può essere influenzata dall'ordine dei casi del file. Può essere utile ottenere più soluzioni diverse con casi disposti in ordini random diversi per verificare la stabilità di una soluzione specifica.

Ipotesi. 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 Prossimità. È inoltre necessario includere nell'analisi tutte le variabili significative. L'omissione di variabili importanti può portare a soluzioni improprie. Poiché l'analisi cluster gerarchica rappresenta un metodo esplorativo, i risultati devono essere considerati provvisori finché non vengano confermati da un campione indipendente.

Per ottenere una Analisi cluster gerarchica

1. Dai menu, scegliere:

   Analizza > **Classifica** > **Cluster gerarchico...**
2. 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.

Analisi cluster gerarchica: Metodo

Metodo cluster. Le alternative disponibili sono: Legame medio fra i gruppi, Legame medio entro gruppi, Del vicino più vicino, Del vicino più lontano, Baricentro, 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 al quadrato, 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 al quadrato, 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 prossimità (non disponibile per i dati binari). I metodi di standardizzazione disponibili sono Punteggi *z*, Intervallo da –1 a 1, Intervallo da 0 a 1, Grandezza massima di 1, Media di 1 e Deviazione standard di 1.

Trasforma misure. Consente di trasformare i valori generati dalla misura della distanza. Questi verranno applicati dopo il calcolo della misura della distanza. Le alternative disponibili sono Valori assoluti, Cambia segno e Modifica scala su intervallo 0-1.

Analisi cluster gerarchica: Statistiche

Pianificazione 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 di prossimità. Fornisce le distanze o le similarità tra gli elementi.

Appartenenza cluster. 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.

Analisi 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.

Stalattite. Visualizza un *grafico a ghiacciolo*, inclusi tutti i cluster o l'intervallo di cluster specificato. Nei grafici a ghiacciolo 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.

Analisi cluster gerarchica: Salva nuove variabili

Appartenenza cluster. 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.

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

Capitolo 26. Analisi del cluster 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 conoscono queste informazioni. 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 cluster iniziali, tabella ANOVA. Ogni caso: informazioni sul cluster, distanza dal centro del cluster.

Considerazioni sui dati relativi all'analisi del cluster delle K Medie

Dati. Le variabili devono essere quantitative a livello di intervallo o di rapporto. Se le variabili sono binarie o conteggi, utilizzare la procedura Analisi cluster gerarchica.

Ordine dei casi e centri cluster iniziali. L'algoritmo predefinito per la scelta dei centri di cluster iniziali varia a seconda dell'ordine dei casi. L'opzione**Usa medie mobili** della finestra di dialogo Itera 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 random 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.

Ipotesi. Le distanze vengono calcolate utilizzando la distanza euclidea semplice. Se si desidera utilizzare un'altra misura di distanza o di similarità, utilizzare la procedura Analisi cluster gerarchica. Lo scaling 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 del cluster delle K Medie

1. Dai menu, scegliere:
   Analizza > **Classifica** > **Cluster delle k medie...**

2. Selezionare le variabili da utilizzare nell'analisi cluster.
3. 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 di dati.
4. Selezionare il metodo **Itera e classifica** oppure il metodo **Classifica soltanto**.
5. In alternativa, selezionare una variabile di identificazione per etichettare i casi.

Efficienza dell'analisi del cluster delle 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 cluster, 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 dataset. I dataset possono anche essere riutilizzati nella stessa sessione, ma non vengono salvati come file a meno che siano stati salvati come tali al termine della sessione. I nomi dei dataset devono essere conformi alle regole di denominazione delle variabili. Per ulteriori informazioni, consultare l'argomento .

Analisi del cluster delle K medie: Itera

Nota: queste opzioni sono disponibili solo se si seleziona il metodo **Itera e classifica** dalla finestra di dialogo Analisi del cluster delle k Medie.

Numero massimo 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.

Analisi del cluster delle K medie: Salva

È possibile salvare informazioni sulla soluzione come nuove variabili da utilizzare in analisi successive:

Appartenenza cluster. 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 del cluster. Consente di creare una nuova variabile che indica la distanza euclidea tra ciascun caso e il relativo centro di classificazione.

Analisi del cluster delle K medie: Opzioni

Statistiche. È possibile selezionare le seguenti statistiche: centri cluster iniziali, tabella ANOVA e informazioni sul cluster per ciascun caso.

- *Centri cluster iniziali*. Prima stima delle medie variabile per ciascun cluster. In mancanza di indicazioni particolari, viene selezionato dai dati un numero casuale 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 che include i test F univariati per ogni variabile di clustering. I test F sono descrittivi e il livello di significatività fornisce informazioni utili. La tabella ANOVA non viene visualizzata se tutti i casi sono assegnati a un solo cluster.
- *Informazioni cluster per ogni caso*. Visualizza per ciascun caso l'assegnazione cluster finale e la distanza euclidea tra il caso e il centro del cluster usato per classificare il caso. Visualizza anche la distanza euclidea tra i centri del cluster finali.

Valori mancati. Le opzioni disponibili sono **Escludi casi a livello di elenco** o **Escludi casi a coppie**.

- **Escludi casi a livello di elenco.** Consente di escludere i casi coni valori mancanti per le variabili di raggruppamento dall'analisi.
- **Escludi casi a coppie.** Consente di assegnare i casi ai cluster in base alle distanze calcolate da tutte le variabili con valori non mancanti.

Funzioni 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 il passaggio di dati normalmente usato per stimarli.
- Specificare i centri iniziali dei cluster direttamente come parte della sintassi del comando.
- Specificare i nomi delle variabili salvate.

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

Capitolo 27. Test non parametrici

I test non parametrici formulano ipotesi 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 di 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.

Qual è il proprio l'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 del 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 di esecuzione 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 osservi che questa impostazione viene selezionata automaticamente se successivamente si apportano delle modifiche alle opzioni nella scheda Impostazioni che sono incompatibili con l'obiettivo attualmente selezionato.

Per ottenere test non parametrici a campione singolo

Dai menu, scegliere:

Analizza > **Test non parametrici** > **Campione singolo...**

1. 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

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 test. È obbligatorio avere almeno un campo test.

Utilizza assegnazioni campi personalizzate. Questa opzione consente di ignorare i ruoli dei campi. Dopo averla selezionata, compilare i campi riportati sotto:

- **Campi test.** Selezionare uno o più campi.

Scheda Impostazioni

La scheda Impostazioni contiene vari gruppi di impostazioni differenti che è possibile modificare per definire il modo in cui l'algoritmo elabora i dati. Se si apportano delle modifiche alle impostazioni predefinite che sono incompatibili con l'obiettivo attualmente selezionato, la scheda Obiettivo viene aggiornata automaticamente per selezionare l'opzione **Personalizza analisi**.

Scegli test

Queste impostazioni specificano i test da eseguire sui campi indicati nella scheda Campi.

Scegli automaticamente i test in base ai dati. Questa impostazione applica il test binomiale ai campi categoriali con due sole categorie valide (non mancanti), il test del 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 indicatore (campo categoriale con due sole categorie) è uguale a quella prevista da una distribuzione binomiale specificata. È possibile inoltre richiedere gli intervalli di confidenza. Consultare “Test binomiale: Opzioni” per i dettagli sulle impostazioni del test.
- **Confronta le probabilità osservate con quelle ipotizzate (test chi-quadrato).** Il test del 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. Consultare “Test del chi-quadrato: Opzioni” a pagina 129 per i dettagli sulle impostazioni del test.
- **Prova la distribuzione osservata con quella ipotizzata (test di Kolmogorov-Smirnov).** Il test di Kolmogorov-Smirnov viene applicato ai campi continui e ordinali. Questa opzione genera un test a campione singolo che verifica se la funzione di distribuzione cumulativa del campione di un campo è omogenea con una distribuzione uniforme, normale, di Poisson o esponenziale. Consultare “Opzioni test di Kolmogorov-Smirnov” a pagina 129 per i dettagli sulle impostazioni del test.
- **Confronta mediana osservata con quella ipotizzata (test dei ranghi con segni di Wilcoxon).** Il test dei ranghi con segni di Wilcoxon si applica ai campi continui e ordinali. Questa opzione genera un test a campione singolo del valore mediano di un campo. Specificare un numero come mediana ipotizzata.
- **Prova la casualità della sequenza (test di esecuzione).** Il test di esecuzione 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 random. Consultare “Opzioni test di esecuzione” a pagina 129 per i dettagli sulle impostazioni del test.

Test binomiale: Opzioni: Il test binomiale è destinato ai campi indicatore (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 dei record definiti come "esiti positivi", o *p*. Specificare un valore maggiore di 0 e minore di 1. Il valore predefinito è 0.5.

Intervallo di confidenza. Sono disponibili i seguenti metodi per calcolare gli intervalli di confidenza per i dati binari.

- **Clopper-Pearson (esatto).** Un intervallo esatto basato sulla distribuzione binomiale cumulativa.
- **Jeffreys.** Un intervallo bayesiano basato sulla distribuzione a posteriori di *p* utilizzando la probabilità 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 predefinita.
- **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 divisione.

- **Valore intermedio campione** imposta il punto di divisione sulla media dei valori massimo e minimo.
- **Punto di divisione personalizzato** consente di specificare un valore per il punto di divisione.

Test del chi-quadrato: Opzioni: Tutte le categorie hanno uguale probabilità. Genera frequenze uguali fra tutte le categorie del campione. È l'impostazione predefinita.

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 specificano le probabilità previste personalizzate, i valori delle categorie personalizzate devono includere tutti i valori dei campi nei dati; altrimenti, il test non viene eseguito per quel campo.

Opzioni test di Kolmogorov-Smirnov: Questa finestra di dialogo specifica le distribuzioni da testare e i parametri delle distribuzioni ipotizzate.

Normale. Utilizzare dati campione utilizza la media osservata e la deviazione standard, **Personalizzato** consente di specificare i valori.

Uniforme. Utilizza dati campione utilizza il minimo e il massimo osservati, **Personalizzato** consente di specificare dei valori.

Esponenziale. Media del campione utilizza la media osservata, **Personalizzato** consente di specificare i valori.

Poisson. Media del campione utilizza la media osservata, **Personalizzato** consente di specificare dei valori.

Opzioni test di esecuzione: Il test di esecuzione è destinato ai campi indicatore (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. Sono disponibili le seguenti opzioni:

- **Il campione contiene solo 2 categorie** esegue il test di esecuzione 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 di esecuzione 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 divisione per i campi continui. Specifica come vengono definiti i gruppi per i campi continui. Il primo gruppo è definito come valori uguali o inferiori a un punto di divisione.

- **Mediana campione** imposta il punto di divisione sulla mediana del campione.
- **Media del campione** imposta il punto di divisione sulla media del campione.
- **Personalizzato** consente di specificare un valore per il punto di divisione.

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. Specificare un valore numerico compreso tra 0 e 100. Il valore predefinito è 95.

Casi esclusi. Specifica come determinare la base di casi per i test.

- **Escludi casi a livello di elenco** significa che i record con valori mancanti in qualunque campo denominato nella scheda Campi vengono esclusi da tutte le analisi.
- **Escludi 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

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.

Funzioni 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.

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

Test non parametrici di campioni indipendenti

I test non parametrici di 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.

Qual è il proprio l'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 osservi che questa impostazione viene selezionata automaticamente se successivamente si apportano delle modifiche alle opzioni nella scheda Impostazioni che sono incompatibili con l'obiettivo attualmente selezionato.

Per ottenere test non parametrici di campioni indipendenti

Dai menu, scegliere:

Analizza > **Test non parametrici** > **Campioni indipendenti...**

1. 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

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 e ordinali con un ruolo predefinito come Obiettivo o Entrambi saranno utilizzati come campi test. Se vi è un solo campo categoriale con un ruolo predefinito come Input, verrà utilizzato come un campo di raggruppamento. Altrimenti, per impostazione predefinita non viene utilizzato alcun campo di raggruppamento ed è necessario utilizzare le assegnazioni campi personalizzate. È obbligatorio disporre di almeno un campo 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 test.** Selezionare uno o più campi continui od ordinali.
- **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 delle modifiche alle impostazioni predefinite che sono incompatibili con l'obiettivo attualmente selezionato, la scheda Obiettivo viene aggiornata automaticamente per selezionare l'opzione **Personalizza analisi**.

Scegli test

Queste impostazioni specificano i test da eseguire sui campi indicati nella scheda Campi.

Scegli automaticamente i test in base ai dati. 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 di 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 di esecuzione 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 comparazioni multiple dei *k* campioni, ovvero comparazioni multiple **tutto a coppie** o comparazioni **decrescenti a fasi**.

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 la grandezza 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 ubicazione 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 ubicazione 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 comparazioni multiple dei *k* campioni, ovvero comparazioni multiple di tipo **Tutto a coppie** o comparazioni di tipo **Decrescente a fasi**.

- **Confronta gli intervalli nei gruppi.** Questa opzione genera un test di 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 comparazione. Il primo valore in ordine crescente del campo di raggruppamento definisce il gruppo di controllo, mentre il secondo definisce il gruppo di comparazione. 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 di 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 del dataset), sia un valore personalizzato. Se lo si desidera, è possibile richiedere comparazioni multiple dei *k* campioni, ovvero comparazioni multiple di tipo **Tutto a coppie** o comparazioni di tipo **Decrescente a fasi**.
- **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

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. Specificare un valore numerico compreso tra 0 e 100. Il valore predefinito è 95.

Casi esclusi. Specifica come determinare la base di casi per i test. **Escludi casi a livello di elenco** significa che i record con valori mancanti in qualunque campo denominato in un sottocomando vengono esclusi da tutte le analisi. **Escludi 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

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.

Funzioni 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.

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

Test non parametrici di campioni correlati

Identificare le differenze tra 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 il quale due o più misure correlate vengono memorizzate in campi separati del dataset. Ad esempio, uno studio relativo all'efficacia di un piano di dieta può essere analizzato utilizzando i test non parametrici a campioni correlati, se il peso di ciascun soggetto viene misurato a intervalli regolari e memorizzato in campi quali *Peso prima della dieta*, *Peso intermedio* e *Peso dopo la dieta*. Questi campi sono "correlati".

Qual è il proprio l'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 dei ranghi con 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 osservi che questa impostazione viene selezionata automaticamente se successivamente si apportano delle modifiche alle opzioni nella scheda Impostazioni che sono incompatibili con l'obiettivo attualmente selezionato.

Quando si specificano campi con livelli di misurazione diversi, essi vengono prima separati dal livello di misurazione, quindi si applica il test appropriato a ciascun gruppo. Ad esempio, se si sceglie **Confronta automaticamente i dati osservati con quelli ipotizzati** come obiettivo e si specificano 3 campi continui e 2 campi nominali, il test di Friedman viene applicato ai campi continui e il test di McNemar viene applicato ai campi nominali.

Per ottenere test non parametrici di campioni correlati

Dai menu, scegliere:

Analizza > **Test non parametrici** > **Campioni correlati...**

1. 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

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 test. È obbligatorio avere almeno due campi test.

Utilizza assegnazioni campi personalizzate. Questa opzione consente di ignorare i ruoli dei campi. Dopo averla selezionata, compilare i campi riportati sotto:

- **Campi 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 il modo in cui la procedura elabora i dati. 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

Queste impostazioni specificano i test da eseguire sui campi indicati nella scheda Campi.

Scegli automaticamente i test in base ai dati. 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 dei ranghi con 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.** Il **test di McNemar (2 campioni)** può essere applicato ai campi categoriali. Questa opzione genera un test a campioni correlati per verificare se le combinazioni di valori tra due campi indicatore (campi categoriali con due soli valori) hanno uguale probabilità. Se nella scheda Campi sono specificati più di due campi, il test non viene eseguito. Consultare "Test di McNemar: definisci esito positivo" per i dettagli sulle impostazioni del 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 indicatore (campi categoriali con due soli valori) hanno uguale probabilità. Se lo si desidera è possibile richiedere comparazioni multiple dei *k* campioni, ovvero comparazioni multiple **tutto a coppie** o comparazioni **decrescenti a fasi**. Consultare "Q di Cochran: definisci esito positivo" a pagina 135 per i dettagli sulle impostazioni del test.
- **Test per il cambiamento nei dati multinomiali.** Il **test di omogeneità marginale (2 campioni)** produce un test di campioni correlati che indica se le combinazioni di valori tra due campi ordinali accoppiati sono probabili allo stesso modo. 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 produce un test di campioni correlati per verificare se la differenza mediana tra due campi è diversa da 0. Il test si applica ai campi continui e ordinali. Se nella scheda Campi sono specificati più di due campi, questi 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 accoppiati. Il test si applica ai campi continui ed ordinali. Se nella scheda Campi sono specificati più di due campi, questo 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 il punteggio di vari elementi (campi) da parte di un giudice. Se lo si desidera, è possibile richiedere comparazioni multiple dei *k* campioni, ovvero comparazioni multiple di tipo **Tutto a coppie** o comparazioni di tipo **Decrescente a fasi**.
- **Compara distribuzioni. Test ANOVA per ranghi a due vie di Friedman (k campioni)** produce un test di campioni correlati che indica se i *k* campioni correlati derivano dalla stessa popolazione. Se lo si desidera, è possibile richiedere comparazioni multiple dei *k* campioni, ovvero comparazioni multiple di tipo **Tutto a coppie** o comparazioni di tipo **Decrescente a fasi**.

Test di McNemar: definisci esito positivo: Il test di McNemar è destinato ai campi indicatore (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: Il test Q di Cochran è destinato ai campi indicatore (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

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. Specificare un valore numerico compreso tra 0 e 100. Il valore predefinito è 95.

Casi esclusi. Specifica come determinare la base di casi per i test.

- **Escludi casi a livello di elenco** significa che i record con valori mancanti in qualunque campo denominato in un sottocomando vengono esclusi da tutte le analisi.
- **Escludi 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

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.

Funzioni 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.

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

Vista Modello

Vista Modello

La procedura crea un oggetto Visualizzatore modelli nel visualizzatore. Attivando l'oggetto con un doppio clic, si accede a una vista interattiva del modello La vista Modello ha una finestra composta da due pannelli, la vista principale sul lato sinistro e la vista collegata, o ausiliaria, sul lato destro.

Vi sono due viste principali:

- Riepilogo ipotesi. Questa è la vista predefinita. Per ulteriori informazioni, consultare l'argomento "Riepilogo ipotesi".
- Riepilogo intervallo di confidenza. Per ulteriori informazioni, consultare l'argomento "Riepilogo intervallo di confidenza".

Vi sono sette viste collegate/ausiliarie:

- Test campione singolo. Questa è la vista predefinita se sono stati richiesti i test a un campione. Per ulteriori informazioni, consultare l'argomento "Test campione singolo".
- Test campioni correlati. Questa è la vista predefinita se sono stati richiesti i test a campioni correlati e nessun test a un campione. Per ulteriori informazioni, consultare l'argomento "Test campioni correlati" a pagina 137.
- Test campioni indipendenti. Questa è la vista predefinita se non è stato richiesto alcun test a campioni correlati o test a un campione. Per ulteriori informazioni, consultare l'argomento "Test campioni indipendenti" a pagina 139.
- Informazioni sul campo categoriale. Per ulteriori informazioni, consultare l'argomento "Informazioni sul campo categoriale" a pagina 140.
- Informazioni sul campo continuo. Per ulteriori informazioni, consultare l'argomento "Informazioni sul campo continuo" a pagina 140.
- Comparazioni a coppie. Per ulteriori informazioni, consultare l'argomento "Comparazioni a coppie" a pagina 140.
- Sottoinsiemi omogenei. Per ulteriori informazioni, consultare l'argomento "Sottoinsiemi omogenei" a pagina 140.

Riepilogo ipotesi

La visualizzazione Riepilogo del modello è un'istantanea, un riepilogo immediato dei test non parametrici. Enfatizza le ipotesi e le decisioni null, ponendo l'attenzione sui valori *p* significativi.

- Ciascuna riga corrisponde a un test separato. Facendo clic su una riga, vengono mostrate altre informazioni sul test nella vista collegata.
- Facendo clic sull'intestazione di una colonna, le righe vengono ordinate in base ai valori in quella colonna.
- Il pulsante **Reimposta** consente di ripristinare lo stato originale del Visualizzatore modelli.
- L'elenco a discesa **Filtro del campo** consente di visualizzare solo i test che interessano il campo selezionato.

Riepilogo intervallo di confidenza

Il Riepilogo intervallo di confidenza mostra gli intervalli di confidenza generati dai test non parametrici.

- Ogni riga corrisponde a un intervallo di confidenza separato.
- Facendo clic sull'intestazione di una colonna, le righe vengono ordinate in base ai valori in quella colonna.

Test campione singolo

La vista Test campione singolo mostra i dettagli relativi ai test non parametrici a campione singolo richiesti. Le informazioni mostrate dipendono dal test selezionato.

- L'elenco a discesa **Test** consente di selezionare un determinato tipo di test a campione singolo.
- L'elenco a discesa **Campo(i)** consente di selezionare un campo che è stato testato utilizzando il test selezionato nell'elenco a discesa **Test**.

Test binomiale

Il Test binomiale mostra un grafico a barre in pila e una tabella Test.

- Il grafico a barre in pila visualizza le frequenze osservate e ipotizzate per le categorie "esito positivo" ed "errore" del campo Test, in cui gli "errori" sono impilati sopra gli "esiti positivi". Passando il cursore del mouse su una barra, vengono mostrate le percentuali della categoria in un suggerimento. Le differenze visibili nelle barre indicano che il campo Testo non può avere la distribuzione binomiale ipotizzata.
- La tabella mostra i dettagli del test.

Test del chi-quadrato

La vista Test del chi-quadrato mostra un grafico a barre raggruppate e una tabella Test.

- Il grafico a barre raggruppate visualizza le frequenze osservate e ipotizzate per ogni categoria del campo Test. Passando il cursore del mouse su una barra, vengono mostrate le frequenze osservate e ipotizzate e la relativa differenza (residuo) in un suggerimento. Le differenze visibili nelle barre osservate rispetto a quelle ipotizzate indicano che il campo Testo non può avere la distribuzione ipotizzata.
- La tabella mostra i dettagli del test.

Ranghi con segni di Wilcoxon

La vista Test dei ranghi con segni di Wilcoxon mostra un istogramma e una tabella Test.

- L'istogramma include delle linee verticali che mostrano le mediane osservate e ipotizzate.
- La tabella mostra i dettagli del test.

Test di esecuzione

La vista Test di esecuzione mostra un grafico a barre e una tabella Test.

- Il grafico visualizza una distribuzione normale con il numero di esecuzioni osservate contrassegnate con una linea verticale. Si osservi che quando viene eseguito il test esatto, il test non si basa sulla distribuzione normale.
- La tabella mostra i dettagli del test.

Test di Kolmogorov-Smirnov

La vista Test di Kolmogorov-Smirnov mostra un istogramma e una tabella Test.

- L'istogramma include una sovrapposizione della funzione di densità di probabilità per la distribuzione esponenziale, Poisson, normale o uniforme ipotizzata. Si osservi che il test si basa sulle distribuzioni cumulative e le differenze più estreme riportate nella tabella devono essere interpretate rispetto alle distribuzioni cumulative.
- La tabella mostra i dettagli del test.

Test campioni correlati

La vista Test campione singolo mostra i dettagli relativi ai test non parametrici a campione singolo richiesti. Le informazioni mostrate dipendono dal test selezionato.

- L'elenco a discesa **Test** consente di selezionare un determinato tipo di test a campione singolo.
- L'elenco a discesa **Campo(i)** consente di selezionare un campo che è stato testato utilizzando il test selezionato nell'elenco a discesa **Test**.

Test di McNemar

La vista Test di McNemar mostra un grafico a barre raggruppate e una tabella Test.

- Il grafico a barre raggruppate visualizza le frequenze osservate e ipotizzate per le celle esterne alla diagonale della tabella 2×2 definita dai campi Test.
- La tabella mostra i dettagli del test.

Test dei segni

La vista Test dei segni mostra un istogramma in pila e una tabella Test.

- L'istogramma in pila visualizza le differenze tra i campi, utilizzando il segno della differenza come campo di sovrapposizione.
- La tabella mostra i dettagli del test.

Test dei ranghi con segni di Wilcoxon

La vista Test dei ranghi con segni di Wilcoxon mostra un istogramma in pila e una tabella Test.

- L'istogramma in pila visualizza 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

La vista Test di omogeneità marginale mostra un grafico a barre raggruppate e una tabella Test.

- Il grafico a barre raggruppate visualizza le frequenze osservate per le celle esterne alla diagonale della tabella definita dai campi Test.
- La tabella mostra i dettagli del test.

Test Q di Cochran

La vista Test Q di Cochran mostra un grafico a barre in pila e una tabella Test.

- Il grafico a barre in pila visualizza le frequenze osservate per le categorie "esito positivo" ed "errore" dei campi di test, in cui gli "errori" sono impilati sopra gli "esiti positivi". Passando il cursore del mouse su una barra, vengono mostrate le percentuali della categoria in un suggerimento.
- La tabella mostra i dettagli del test.

Analisi della varianza per ranghi a due vie di Friedman

La vista Analisi della varianza per ranghi a due vie di Friedman mostra gli istogrammi a pannelli e una tabella Test.

- Gli istogrammi visualizzano la distribuzione osservata delle classificazioni, suddivisa in pannelli dai campi Test.
- La tabella mostra i dettagli del test.

Coefficiente di concordanza di Kendall

La vista Coefficiente di concordanza di Kendall mostra gli istogrammi a pannelli e una tabella Test.

- Gli istogrammi visualizzano la distribuzione osservata delle classificazioni, suddivisa in pannelli dai campi Test.
- La tabella mostra i dettagli del test.

Test campioni indipendenti

La vista Test campioni indipendenti mostra i dettagli relativi ai test non parametrici dei campioni indipendenti richiesti. Le informazioni mostrate dipendono dal test selezionato.

- L'elenco a discesa **Test** consente di selezionare un determinato tipo di test di campioni indipendenti.
- L'elenco a discesa **Campo(i)** consente di selezionare un test e una combinazione di campi di raggruppamenti che è stata testata utilizzando il test selezionato nell'elenco a discesa **Test**.

Test di Mann-Whitney

La vista Test di Mann-Whitney mostra un grafico a piramide della popolazione e una tabella Test.

- Il grafico a piramide della popolazione visualizza gli istogrammi in sequenza per categorie di campo di raggruppamento, rilevando il numero di record in ciascun gruppo e la classificazione della media del gruppo.
- La tabella mostra i dettagli del test.

Test di Kolmogorov-Smirnov

La vista Test di Kolmogorov-Smirnov mostra un grafico a piramide della popolazione e una tabella Test.

- Il grafico a piramide della popolazione visualizza gli istogrammi in sequenza per categorie di campo di raggruppamento, rilevando il numero di record in ciascun gruppo. Le linee di distribuzione cumulative osservate possono essere visualizzate o nascoste facendo clic sul pulsante **Cumulativo**.
- La tabella mostra i dettagli del test.

Test di esecuzione di Wald-Wolfowitz

La vista Test di esecuzione di Wald-Wolfowitz mostra un grafico a barre in pila e una tabella Test.

- Il grafico a piramide della popolazione visualizza gli istogrammi in sequenza per categorie di campo di raggruppamento, rilevando il numero di record in ciascun gruppo.
- La tabella mostra i dettagli del test.

Test di Kruskal-Wallis

La vista Test di Kruskal-Wallis mostra i grafici a scatole e una tabella Test.

- Per ogni categoria del campo di raggruppamento vengono visualizzati dei grafici a scatole separati. Passando il cursore del mouse su una casella, viene mostrata la classificazione della media in un suggerimento.
- La tabella mostra i dettagli del test.

Test di Jonckheere-Terpstra

La vista Test di Jonckheere-Terpstra mostra i grafici a scatole e una tabella Test.

- Per ogni categoria del campo di raggruppamento vengono visualizzati dei grafici a scatole separati.
- La tabella mostra i dettagli del test.

Test delle reazioni estreme di Moses

La vista Test delle reazioni estreme di Moses mostra i grafici a scatole e una tabella Test.

- Per ogni categoria del campo di raggruppamento vengono visualizzati dei grafici a scatole separati. Le etichette dei punti possono essere visualizzate o nascoste facendo clic sul pulsante **ID record**.
- La tabella mostra i dettagli del test.

Test della mediana

La vista Test della mediana mostra i grafici a scatole e una tabella Test.

- Per ogni categoria del campo di raggruppamento vengono visualizzati dei grafici a scatole separati.
- La tabella mostra i dettagli del test.

Informazioni sul campo categoriale

La vista Informazioni sul campo categoriale visualizza un grafico a barre per il campo categoriale selezionato nell'elenco a discesa **Campo(i)**. L'elenco di campi disponibili è limitato ai campi categoriali utilizzati nel test selezionato attualmente nella vista Riepilogo ipotesi.

- Passando il cursore del mouse su una barra, vengono mostrate le percentuali della categoria in un suggerimento.

Informazioni sul campo continuo

La vista Informazioni sul campo continuo visualizza un istogramma per il campo continuo selezionato nell'elenco a discesa **Campo(i)**. L'elenco di campi disponibili è limitato ai campi continui utilizzati nel test selezionato attualmente nella vista Riepilogo ipotesi.

Comparazioni a coppie

La vista Comparazioni a coppie mostra un grafico di rete della distanza e la tabella di comparazioni prodotta dai test non parametrici di *k* campioni quando sono richieste più comparazioni a coppie.

- Il grafico di rete della distanza è una rappresentazione grafica della tabella di comparazioni in cui le distanze tra i nodi nella rete corrispondono alle differenze tra i campioni. Le linee gialle corrispondono alle differenze significative dal punto di vista statistico; le linee nere corrispondono alle differenze non significative. Passando il cursore del mouse su una linea nella rete viene visualizzato un suggerimento con la significatività adattata della differenza tra i nodi connessi dalla linea.
- La tabella di comparazioni mostra i risultati numerici di tutte le comparazioni a coppie. Ogni riga corrisponde a una comparazione a coppie differente. Facendo clic sull'intestazione di una colonna, le righe vengono ordinate in base ai valori in quella colonna.

Sottoinsiemi omogenei

La vista Sottoinsiemi omogenei mostra una tabella di comparazioni prodotta dai test non parametrici di *k* campioni quando vengono richieste più comparazioni decrescenti a fasi.

- Ogni riga nel gruppo Campione corrisponde a un campione correlato separato (rappresentato nei dati da campi separati). I campioni che dal punto di vista statistico non sono molto diversi vengono raggruppati in sottoinsiemi dello stesso colore; esiste una colonna separata per ciascun sottoinsieme identificato. Quando tutti i campioni dal punto di vista statistico sono molto diversi, esiste un sottoinsieme separato per ogni campione. Quando nessuno dei campioni dal punto di vista statistico è molto diverso, esiste un solo sottoinsieme.
- Per ogni sottoinsieme contenente più di un campione vengono calcolati la statistica del test, il valore di significatività e il valore di significatività adattato.

Funzioni 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.

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

Finestre legacy

Esistono anche numerose finestre di dialogo "legacy" che eseguono test non parametrici. Queste finestre di dialogo supportano la funzionalità fornita dall'opzione Test esatti.

Test del chi-quadrato. Consente di ordinare in tabelle una variabile in categorie e di ottenere una statistica 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 di esecuzione. Consente di verificare se l'ordine di occorrenza di due valori di una variabile è random.

Test di Kolmogorov-Smirnov per un campione. Consente di confrontare la funzione di distribuzione cumulativa osservata per una variabile con la distribuzione teorica specificata, che può essere normale, uniforme, esponenziale o di Poisson.

Test di 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 di esecuzione di Wald-Wolfowitz.

Test per due campioni correlati. Consente di confrontare le distribuzioni di due variabili. Sono disponibili il test dei ranghi con segni 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 del chi-quadrato

La procedura Test del chi-quadrato permette di ordinare in tabelle 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 del 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.

Considerazioni sui dati di test del 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.

Ipotesi. I test non parametrici non richiedono ipotesi relative alla forma della distribuzione sottostante. Si presume che i dati rappresentino un campione random. 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 del chi-quadrato

1. Dai menu, scegliere:

   Analizza > **Test non parametrici** > **Finestre di dialogo legacy** > **Chi-quadrato...**
2. Selezionare una o più variabili per il test. Ogni variabile produce un test distinto.
3. 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 del 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 **Usa intervallo specificato** e inserire valori interi per il limite inferiore e superiore. Le categorie verranno definite per ogni valore intero incluso 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 del chi-quadrato verranno utilizzati solo i valori interi compresi tra 1 e 4.

Valori previsti. 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 dell'elenco corrisponde al valore di gruppo minore della variabile, mentre l'ultimo corrisponde al valore maggiore. Gli elementi dell'elenco 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, un elenco valori formato da 3, 4, 5, 4 specifica le proporzioni attese 3/16, 4/16, 5/16 e 4/16.

Test del chi-quadrato: Opzioni

Statistiche. È possibile scegliere una o entrambe le statistiche di riepilogo.

- **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 mancati. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Escludi casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Escludi casi a livello di elenco.** I casi con valori mancanti per qualsiasi variabile sono esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TESTS (test del 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`).

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

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 test 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 una monetina, la probabilità che esca testa è uguale a 1/2. In base a questa ipotesi, una monetina viene lanciata 40 volte e i risultati vengono registrati (testa o croce). 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.

Considerazioni sui dati di test binomiale

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 può avere solo due valori possibili: *sì* o *no*, *true* o *false*, 0 o 1, e così via. Il primo valore rilevato nel dataset 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.

Ipotesi. I test non parametrici non richiedono ipotesi relative alla forma della distribuzione sottostante. Si presume che i dati rappresentino un campione random.

Per ottenere un test binomiale

1. Dai menu, scegliere:

   Analizza > **Test non parametrici** > **Finestre di dialogo legacy** > **Binomiale...**
2. Selezionare una o più variabili numeriche oggetto del test.
3. È possibile fare clic su **Opzioni** per ottenere statistiche descrittive, quartili e controllo delle modalità di elaborazione dei dati mancanti.

Test binomiale: Opzioni

Statistiche. È possibile scegliere una o entrambe le statistiche di riepilogo.

- **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 mancati. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Escludi casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Escludi casi a livello di elenco.** 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`).

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

Test di esecuzione

Il test di esecuzione verifica se l'ordine delle occorrenze di due valori di una variabile è random. Una successione è una sequenza di osservazioni simili. Un campione con troppe o troppo poche successioni indica che il campione non è random.

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 di esecuzione per determinare se il campione è stato estratto in modo random.

Statistiche. Media, deviazione standard, valore minimo, valore massimo, numero di casi non mancanti e quartili.

Considerazioni sui dati di test di esecuzione

Dati. Le variabili devono essere numeriche. Per convertire le variabili stringa in variabili numeriche, utilizzare il comando Ricodifica automatica del menu Trasforma.

Ipotesi. I test non parametrici non richiedono ipotesi relative alla forma della distribuzione sottostante. Utilizzare campioni da distribuzioni di probabilità continue.

Per ottenere un test di esecuzione

1. Dai menu, scegliere:

   Analizza > **Test non parametrici** > **Finestre di dialogo legacy** > **Esecuzioni...**
2. Selezionare una o più variabili numeriche oggetto del test.
3. È possibile fare clic su **Opzioni** per ottenere statistiche descrittive, quartili e controllo delle modalità di elaborazione dei dati mancanti.

Test di esecuzione: 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 di esecuzione

Statistiche. È possibile scegliere una o entrambe le statistiche di riepilogo.

- **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 mancati. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Escludi casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Escludi casi a livello di elenco.** I casi con valori mancanti per qualsiasi variabile sono esclusi da tutte le analisi.

Funzioni aggiuntive del comando NPAR TESTS (Test di esecuzione)

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`).

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

Test di Kolmogorov-Smirnov per un campione

La procedura Test di Kolmogorov-Smirnov per un campione consente di confrontare la funzione di distribuzione cumulativa 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 cumulativa 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.

Considerazioni sui dati di test di Kolmogorov-Smirnov per un campione

Dati. Utilizzare variabili quantitative (misurazione a livello di intervallo o di rapporto).

Ipotesi. 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 adattato (selezionabile dalla procedura Esplora).

Per ottenere un test di Kolmogorov-Smirnov per un campione

1. Dai menu, scegliere:
   Analizza > **Test non parametrici** > **Finestre di dialogo legacy** > **K-S per 1 campione...**
2. Selezionare una o più variabili numeriche oggetto del test. Ogni variabile produce un test distinto.
3. 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

Statistiche. È possibile scegliere una o entrambe le statistiche di riepilogo.

- **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 mancati. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Escludi casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Escludi casi a livello di elenco.** 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`).

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

Test di 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: *U* di Mann-Whitney, Reazioni estreme di Moses, *Z* di Kolmogorov-Smirnov, test di esecuzione di Wald-Wolfowitz.

Considerazioni sui dati per i test per due campioni indipendenti

Dati. Utilizzare variabili numeriche che possono essere ordinate.

Ipotesi. Utilizzare campioni random 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

1. Dai menu, scegliere:

   Analizza > **Test non parametrici** > **Finestre di dialogo legacy** > **2 campioni indipendenti...**
2. Selezionare una o più variabili numeriche.
3. 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 dell'ubicazione delle due popolazioni campione. Le osservazioni di entrambi i gruppi vengono combinate e ordinate per ranghi, assegnando il rango medio in caso di correlazioni. Il numero di correlazioni deve essere inferiore al numero totale di osservazioni. Se l'ubicazione 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 la classificazione della media minore, a meno che i gruppi non abbiano la stessa classificazione della media: 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 di esecuzione di Wald-Wolfowitz** sono test di carattere più generale che consentono di individuare le differenze tra le distribuzioni in termini di forma e ubicazione. 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 di esecuzione 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 classificazione.

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 ritagliato il 5% dei casi di controllo.

Test per due campioni indipendenti: Definisci gruppi

Per dividere il file in due gruppi o campioni, immettere un valore intero per Gruppo 1 e un altro valore per Gruppo 2. I casi con altri valori vengono esclusi dall'analisi.

Test per due campioni indipendenti: Opzioni

Statistiche. È possibile scegliere una o entrambe le statistiche di riepilogo.

- **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 mancati. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Escludi casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Escludi casi a livello di elenco.** 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 ritagliare dal test di Moses (con il sottocomando `MOSES`).

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

Test per due campioni correlati

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 del rango con segno 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: ranghi con segni di Wilcoxon, del segno, McNemar. Se è installata l'opzione Test esatti (disponibile solo nei sistemi operativi Windows), è disponibile anche il test di omogeneità marginale.

Considerazioni sui dati per un test per due campioni dipendenti

Dati. Utilizzare variabili numeriche che possono essere ordinate.

Ipotesi. 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

1. Dai menu, scegliere:

   Analizza > **Test non parametrici** > **Finestre di dialogo legacy** > **2 campioni correlati...**
2. 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 del rango con segno 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 correlate. Se le due variabili sono distribuite in modo analogo, il numero di differenze positive e negative non differirà in misura significativa. Il **test del rango con segno di Wilcoxon** prende in considerazione le informazioni relative al segno e alla grandezza delle differenze tra le coppie. Poiché il test del rango con segno 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 progettazioni 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 progettazioni sperimentali del tipo 'prima e dopo'. Il test di omogeneità marginale è disponibile solo se è stato installato il modulo Test esatti.

Test per due campioni dipendenti: Opzioni

Statistiche. È possibile scegliere una o entrambe le statistiche di riepilogo.

- **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 mancati. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Escludi casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Escludi casi a livello di elenco.** 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.

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

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'analisi della varianza a una via 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.

Considerazioni sui dati dei test per diversi campioni indipendenti

Dati. Utilizzare variabili numeriche che possono essere ordinate.

Ipotesi. Utilizzare campioni random 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

1. Dai menu, scegliere:
   Analizza > **Test non parametrici** > **Finestre di dialogo legacy** > **K campioni indipendenti...**
2. Selezionare una o più variabili numeriche.
3. 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 della varianza a una via e consente di rilevare le differenze nell'ubicazione di distribuzione. Il **test della mediana**, che è più generale ma non altrettanto potente, consente di rilevare le differenze distribuzionali nell'ubicazione 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 la grandezza 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

Per definire l'intervallo, immettere valori interi 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

Statistiche. È possibile scegliere una o entrambe le statistiche di riepilogo.

- **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 mancati. Consente di controllare la modalità di elaborazione dei valori mancanti.

- **Escludi casi test per test.** Quando vengono specificati più test, in ciascuno verranno valutati separatamente i valori mancanti.
- **Escludi casi a livello di elenco.** 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).

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

Test per diversi campioni correlati

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: di Friedman, *W* di Kendall e *Q* di Cochran.

Considerazioni sui dati relativi ai test per diversi campioni dipendenti

Dati. Utilizzare variabili numeriche che possono essere ordinate.

Ipotesi. I test non parametrici non richiedono ipotesi relative alla forma della distribuzione sottostante. Utilizzare campioni random dipendenti.

Per ottenere i test per diversi campioni dipendenti

1. Dai menu, scegliere:

   Analizza > **Test non parametrici** > **Finestre di dialogo legacy** > **K campioni correlati...**
2. 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 **test di Friedman** è l'equivalente non parametrico di una progettazione di misure ripetute per un campione o ANOVA a due vie con una osservazione per cella. Friedman verifica l'ipotesi null secondo cui *k* variabili correlate provengono dalla stessa popolazione. Per ogni caso le variabili *k* sono classificate da 1 a *k*. La statistica del test si basa su queste classificazioni.

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

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)

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

Capitolo 28. Analisi a risposta multipla

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 delle frequenze. La procedura Risposte multiple: Tabelle 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. Nella prima domanda si legge: cerchiare tutte le compagnie aeree, tra le seguenti, con le quali si è volato almeno una volta negli ultimi sei mesi in questa rotta--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 il mapping 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 di mapping, 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 delle frequenze 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 ordinate in tabelle.

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 descrittiva di variabile per l'insieme a risposta multipla. L'etichetta può essere costituita al massimo da 40 caratteri.

Per definire gli insiemi a risposta multipla

1. Dai menu, scegliere:
   Analizza > **Risposte multiple** > **Definisci insiemi di variabili...**
2. Selezionare due o più variabili.
3. 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.
4. Immettere un nome univoco per ciascun insieme a risposta multipla.
5. 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 delle frequenze 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 di variabile 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 valore della prima variabile del gruppo. Se le categorie mancanti per la prima variabile sono presenti per altre variabili del gruppo, definire un'etichetta valore per le categorie mancanti.

Valori mancati. I casi con valori mancanti vengono esclusi tabella per tabella. È inoltre possibile scegliere una delle seguenti opzioni o entrambe:

- **Escludi casi a livello di elenco all'interno delle dicotomie.** Consente di escludere i casi con valori mancanti per qualsiasi variabile dalla tabulazione 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 tabulazioni del gruppo se almeno una variabile contiene il valore conteggiato.
- **Escludi casi a livello di elenco all'interno delle categorie.** Consente di escludere i casi con valori mancanti per qualsiasi variabile dalla tabulazione 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'indagine è una variabile elementare. Per analizzare un elemento a risposta multipla, è necessario combinare le variabili in uno dei due tipi di insiemi a risposta multipla; un insieme a dicotomie multiple o 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 delle frequenze. Se risulta che nessun rispondente 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 ordinati in tabelle 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 delle frequenze.

Statistiche. Tabelle delle frequenze 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.

Considerazioni sui dati relativi alle frequenze delle risposte multiple

Dati. Utilizzare gli insiemi a risposta multipla.

Ipotesi. 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

1. Dai menu, scegliere:
   Analizza > **Risposte multiple** > **Frequenze...**
2. Selezionare uno o più insiemi a risposta multipla.

Risposte multiple: Tabelle di contingenza

La procedura Risposte multiple: Tabelle di contingenza consente di incrociare insiemi a risposta multipla definiti, variabili elementari o una combinazione di entrambi. È inoltre possibile ottenere le percentuali delle celle in base a casi o risposte, modificare il trattamento dei valori mancanti oppure ottenere tavole di contingenza accoppiate. 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 di variabile 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 valore della prima variabile del gruppo. Se le categorie mancanti per la prima variabile sono presenti per altre variabili del gruppo, definire un'etichetta valore 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 suddividere 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 sondaggio destinato ai passeggeri di una linea aerea vengono chieste ai passeggeri le seguenti informazioni: cerchiare tutte le compagnie aeree, tra le seguenti, con le quali si è volato 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. Tavola di contingenza con conteggi relativi a celle, righe, colonne e totale e percentuali relative a celle, righe, colonne e totale. Le percentuali delle celle possono basarsi sui casi o sulle risposte.

Considerazioni sui dati relativi alle tabelle di contingenza a risposta multipla

Dati. Utilizzare insiemi a risposta multipla oppure variabili categoriali numeriche.

Ipotesi. 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 tabelle di contingenza a risposta multipla

1. Dai menu, scegliere:

   Analizza > **Risposte multiple** > **Tabelle di contingenza...**
2. Selezionare una o più variabili numeriche o insiemi a risposta multipla per ciascuna dimensione della tavola di contingenza.
3. 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 Livelli.

Risposte multiple, tabelle di contingenza: Definisci intervalli delle variabili

È necessario definire gli intervalli dei valori per tutte le variabili elementari nella tavola di contingenza. Specificare il valore di categoria minimo e massimo intero che si desidera ordinare in tabelle. 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, tabelle di contingenza: Opzioni

Percentuali delle celle. I conteggi delle celle vengono visualizzi sempre. 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 delle celle sui casi (o rispondenti). Questa opzione non è disponibile se si seleziona la corrispondenza delle variabili tra insiemi a categorie multiple. È inoltre possibile basare le percentuali delle 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 mancati. È possibile scegliere una delle seguenti opzioni o entrambe:

- **Escludi casi a livello di elenco all'interno delle dicotomie.** Consente di escludere i casi con valori mancanti per qualsiasi variabile dalla tabulazione 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 tabulazioni del gruppo se almeno una variabile contiene il valore conteggiato.
- **Escludi casi a livello di elenco all'interno delle categorie.** Consente di escludere i casi con valori mancanti per qualsiasi variabile dalla tabulazione 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 ordinata in tabelle 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:

Metti in corrispondenza le variabili tra gli insiemi di risposte. 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 delle celle saranno basate sulle risposte e non sui rispondenti. 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 tabelle delle 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 valore (con il sottocomando `FORMAT`).

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

Capitolo 29. Risultati di report

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 della frequenza 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 di riepilogo diverse sono disposte in righe distinte. Sono inoltre disponibili gli elenchi dei casi, che possono includere o meno le statistiche di riepilogo.

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 interruzione) e che includa statistiche di riepilogo (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 di riepilogo e controlla il formato per la visualizzazione delle colonne di dati.

Colonne di interruzione. Elenca le variabili di interruzione facoltative che suddividono il report in più gruppi e consente di gestire le statistiche di riepilogo e i formati per la visualizzazione delle colonne di interruzione. Se sono presenti più variabili di interruzione, per ogni categoria di ciascuna variabile di interruzione verrà creato un gruppo distinto all'interno delle categorie della variabile di interruzione precedente nell'elenco. Le variabili di interruzione devono essere variabili categoriali discrete che suddividono i casi in un numero limitato di categorie significative. I singoli valori di ciascuna variabile di interruzione 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 di riepilogo globali, la visualizzazione dei valori mancanti e la numerazione delle pagine.

Visualizza casi. Consente di visualizzare i valori effettivi (o etichette valore) delle variabili delle colonne di dati per ciascun caso. In tal modo viene creato un elenco che potrebbe risultare molto più lungo di un report di riepilogo.

Anteprima. 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 interruzione, prima di generarlo è necessario ordinare il file di dati in base ai valori delle variabili di interruzione. Se il file di dati è già ordinato in base ai valori delle variabili di interruzione, è possibile ridurre il tempo di elaborazione selezionando questa opzione. Questa opzione risulta particolarmente utile dopo aver visualizzato un'anteprima del report.

Ottenere un report di riepilogo: riepiloghi per righe

1. Dai menu, scegliere:

   Analizza > **Report** > **Report : Riepiloghi per righe...**
2. Selezionare una o più variabili per Colonne dati. Per ogni variabile selezionata verrà generata una colonna nel report.
3. Per i report ordinati e visualizzati in base ai sottogruppi, selezionare una o più variabili per Colonne di interruzione.
4. Per i report con statistiche di riepilogo per i sottogruppi definiti in base alle variabili di interruzione, selezionare la variabile di interruzione nell'elenco Variabili di colonna di interruzione e fare clic su **Riepilogo** nel gruppo Colonne di interruzione per specificare le misure di riepilogo.
5. Per i report con statistiche di riepilogo globali, fare clic su **Riepilogo** per specificare le misure di riepilogo.

Formato delle colonne e di interruzione 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 dei dati o delle etichette valore. L'opzione Formato colonna dati consente di impostare il formato delle colonne dati nella parte destra della pagina del report. L'opzione Formato di interruzione consente di impostare il formato delle colonne di interruzione nella parte sinistra.

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 interruzioni 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 valore 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 dei dati o delle etichette valore definite. I valori dei dati 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: Righe di riepilogo per/Righe di riepilogo finali

Le due finestre di dialogo Report: Righe di riepilogo consentono di impostare la visualizzazione della statistica di riepilogo per i gruppi di interruzione e per l'intero report. Righe di riepilogo consente di impostare la statistica del sottogruppo per ciascuna categoria definita tramite le variabili di interruzione. Righe di riepilogo finali consente di impostare le statistiche globali visualizzate nella parte finale del report.

Le statistiche di riepilogo 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 interruzione

La funzione Opzioni di interruzione consente di impostare la spaziatura e l'impaginazione delle informazioni sui gruppi.

Controllo pagina. Consente di impostare la spaziatura e l'impaginazione per le categorie relative alla variabile di interruzione 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.

Righe vuote prima dei riepiloghi. Consente di impostare il numero delle linee vuote tra le etichette dell'asse della categoria o i dati e le statistiche di riepilogo. Questa funzionalità risulta particolarmente utile per i report combinati che includono sia l'elenco dei singoli casi che le statistiche di riepilogo per i gruppi. In questo tipo di report è possibile inserire una spaziatura tra l'elenco dei casi e le statistiche di riepilogo.

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.

Escludi casi con valori mancanti a livello di elenco . Consente di eliminare dal report i casi con valori mancanti per le variabili del report.

Valori mancanti visualizzati come . Consente di specificare il simbolo che rappresenta i valori mancanti nel file di 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*.

Numero di pagine 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.

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 interruzione. Consente di impostare la visualizzazione delle colonne di interruzione. Se vengono specificate più variabili di interruzione, queste possono trovarsi in colonne diverse oppure nella prima colonna. Se tutte le variabili di interruzione 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.

Righe colonna di dati ed etichette di interruzione. Consente di posizionare le informazioni relative alle colonne dati (valori dei dati e/o statistiche di riepilogo) in relazione alle etichette di interruzione 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.

Se si inseriscono variabili in titoli e in piè di pagina, l'etichetta valore corrente o il valore della variabile verrà visualizzato nel titolo o nel piè di pagina. Nei titoli viene visualizzata l'etichetta valore

corrispondente al valore della variabile all'inizio della pagina. Nei piè di pagina viene visualizzata l'etichetta valore corrispondente al valore della variabile al termine della pagina. Se non sono presenti etichette valore, 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 di 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 di riepilogo vengono visualizzate in colonne distinte.

Esempio. Un'azienda proprietaria di una catena di negozi al dettaglio conserva le registrazioni delle informazioni sui dipendenti, tra cui gli stipendi, le mansioni e il reparto in cui lavora ogni dipendente. È quindi possibile creare un report che includa le statistiche di riepilogo 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 di riepilogo e di impostare il formato di visualizzazione e le statistiche di riepilogo visualizzate per ogni variabile.

Colonne di interruzione. Consente di visualizzare l'elenco delle variabili di interruzione facoltative che suddividono il report in più gruppi e di impostare i formati di visualizzazione delle colonne di interruzione. Se sono presenti più variabili di interruzione, per ogni categoria di ciascuna variabile di interruzione verrà creato un gruppo distinto all'interno delle categorie della variabile di interruzione precedente nell'elenco. Le variabili di interruzione 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.

Anteprima. 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 interruzione, prima di generarlo è necessario ordinare il file di dati in base ai valori delle variabili di interruzione. Se il file di dati è già ordinato in base ai valori delle variabili di interruzione, è possibile ridurre il tempo di elaborazione selezionando questa opzione. Questa opzione risulta particolarmente utile dopo aver visualizzato un'anteprima del report.

Ottenere un report di riepilogo: riepiloghi per colonne

1. Dai menu, scegliere:

   Analizza > **Report** > **Report: Riepiloghi per colonne...**
2. Selezionare una o più variabili per Colonne dati. Per ogni variabile selezionata verrà generata una colonna nel report.
3. Per modificare la misura di riepilogo relativa a una variabile, selezionare la variabile nell'elenco Variabili di colonna di dati e fare clic su **Riepilogo**.
4. Per ottenere più di una misura di riepilogo per una variabile, selezionare la variabile nell'elenco origine e spostarla nell'elenco Variabili di colonna di dati più volte, una per ogni misura di riepilogo desiderata.

5. 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.
6. Per i report ordinati e visualizzati in base ai sottogruppi, selezionare una o più variabili per Colonne di interruzione.

Funzione di riepilogo delle colonne di dati

L'opzione Righe di riepilogo consente di controllare la statistica di riepilogo visualizzata per la variabile della colonna dati selezionata.

Le statistiche di riepilogo 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 di riepilogo generali che riassumono due o più colonne dati.

Le statistiche di riepilogo 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.

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.

1° colonna - 2° colonna. La colonna *totale* rappresenta la differenza delle colonne nell'elenco Colonna di riepilogo. L'elenco Colonna di riepilogo deve contenere esattamente due colonne.

1° colonna / 2° colonna. La colonna *totale* rappresenta il quoziente delle colonne nell'elenco Colonna di riepilogo. L'elenco Colonna di riepilogo deve contenere esattamente due colonne.

% 1° colonna / 2° 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 interruzione per i report di riepilogo per colonne sono le stesse descritte per i report di riepilogo per righe.

Report: Opzioni di interruzione (Riepiloghi per colonne)

La funzione Opzioni di interruzione consente di impostare la visualizzazione, la spaziatura e l'impaginazione per i gruppi.

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 interruzione 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.

Righe 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.

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 riga di riepilogo.
- Inserire righe di riepilogo nelle colonne dati per le variabili diverse dalle variabili della colonna, o per le varie combinazioni (funzioni composte) di funzioni di riepilogo.
- Utilizzare Mediana, Moda, Frequenza e Percentuale come funzioni di riepilogo.
- Controllare in modo più preciso il formato di visualizzazione delle statistiche di riepilogo.
- Inserire linee vuote in corrispondenza di vari punti nei report.
- Inserire linee vuote dopo ogni *n*° caso nei report elencati.

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.

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

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 interclasse possono essere utilizzati per calcolare le stime di affidabilità tra stimatori.

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 di riepilogo degli elementi, correlazioni interelemento e covarianze, stime di affidabilità, tabella ANOVA, coefficienti di correlazione interclasse, T^2 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 interelemento.
- **Suddivisione a metà.** Questo modello suddivide 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 repliche.
- **Parallelo esatto.** Questo modello afferma le ipotesi del modello parallelo e assume inoltre medie uguali degli elementi.

Considerazioni sui dati dell'analisi di affidabilità

Dati. I dati possono essere dicotomici, ordinali oppure intervalli ma devono essere codificati numericamente.

Ipotesi. 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 l'analisi cluster gerarchica per raggruppare le variabili.

Per ottenere l'analisi di affidabilità

1. Dai menu, scegliere:
   Analizza > **Scala** > **Analisi di affidabilità...**
2. Selezionare due o più variabili come potenziali componenti di una scala additiva.
3. Scegliere un modello dall'elenco a discesa Modello.

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 Suddivisione a metà.** Correlazione tra stime dei parametri, affidabilità di Suddivisione a metà di Guttman, affidabilità di Spearman-Brown (lunghezza uguale e diversa) e coefficiente alfa per ogni metà.
- **Modelli 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, varianza comune e true, correlazione interelemento comune stimata, affidabilità stimata e stima di affidabilità non distorta.

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'elemento è eliminato.** Consente di visualizzare statistiche di riepilogo 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 di riepilogo per le medie degli elementi. Vengono visualizzate la media degli elementi più piccola, la più grande e la media, nonché l'intervallo e la varianza standard delle medie degli elementi e il rapporto fra la media degli elementi più grande e quella più piccola.
- *Varianze*. Statistiche di riepilogo per le varianze degli elementi. Vengono riprodotte la varianza degli elementi minima, massima e media, l'intervallo e la varianza delle varianze degli elementi e il rapporto fra la varianza degli elementi più grande e quella più piccola.
- *Covarianze*. Statistiche di riepilogo per le covarianze interelemento. Vengono visualizzate le covarianze interelemento più piccola, più grande e media, l'intervallo e la varianza delle covarianze interelemento e il rapporto tra le covarianze interelemento più grandi e più piccole.
- *Correlazioni*. Statistiche di riepilogo per le correlazioni interelemento. Vengono visualizzate le correlazioni interelemento più piccola, più grande e media, l'intervallo e la varianza delle correlazioni interelemento e il rapporto tra le correlazioni interelemento più grandi e più piccole.

Interelemento. 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 del chi-quadrato sostituisce il test F solitamente usato nella tabella ANOVA.
- *Chi-quadrato di Cochran*. Visualizza la Q di Cochran. Questa opzione è appropriata per i dati dicotomici. La statistica Q sostituisce la statistica F solitamente usata nella tabella ANOVA.

T-quadrato di Hotelling. Crea un test multivariato dell'ipotesi null in base alla quale tutti gli elementi sulla scala hanno la stessa media.

Test di additività di Tukey. Produce un test dell'ipotesi che non vi sia alcuna 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 random e A una via random. Selezionare **A due vie misto** quando gli effetti relativi alle persone sono random e quelli relativi all'elemento sono fissi, **A due vie random** quando sia gli effetti relativi alle persone che quelli relativi all'elemento sono random oppure **A una via random** quando gli effetti relativi alle persone sono casuali.
- **Tipo.** Consente di selezionare il tipo di indice. I tipi disponibili sono Uniformità e Accordo assoluto.
- **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 sull'ipotesi. Si tratta del valore rispetto al quale viene confrontato il valore osservato. Il valore predefinito è 0.

Funzioni 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 suddivisioni diverse dalle metà uguali per il metodo di suddivisione a metà.

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

Capitolo 31. Scaling multidimensionale

La procedura Scaling multidimensionale consente di effettuare un tentativo per trovare la struttura in un insieme di misure della distanza tra oggetti o casi. Questa attività viene compiuta assegnando le osservazioni a ubicazioni 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 ai punteggi soggettivi 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 punteggi di 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 ogni modello: matrice dei dati, matrice dei dati scalati in modo ottimale, S-stress (di Young), stress (di Kruskal), RSQ, coordinate degli stimoli, stress medio e RSQ per ogni stimolo (modelli RMDS). Per i modelli delle differenze singole (INDSCAL): pesi del soggetto e indice di stranezza per ogni soggetto. Per ogni matrice nei modelli di scaling multidimensionale replicati: stress e RSQ per ogni stimolo. Grafici: coordinate degli stimoli (bi- o tri-dimensionali), grafico a dispersione delle disparità rispetto alle distanze.

Considerazioni sui dati dello Scaling multidimensionale

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.

Ipotesi. 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

1. Dai menu, scegliere:
   Analizza > **Scala** > **Scaling multidimensionale...**
2. Selezionare almeno quattro variabili numeriche per l'analisi.
3. Nel gruppo Distanze selezionare **I dati sono distanze** o **Crea le distanze dai dati**.

4. Se si seleziona **Crea le distanze dai dati**, è anche possibile selezionare una variabile di raggruppamento per le singole matrici. La variabile di raggruppamento può essere di tipo numerico o stringa.

In alternativa, è possibile anche:

- Specificare una forma della matrice di distanza quando i dati sono distanze.
- Specificare la misura di distanza da utilizzare durante la creazione delle distanze dai dati.

Scaling multidimensionale: Forma dei dati

Se il dataset 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 condizionalità della riga.

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 al quadrato, Chebychev, City-Block, Minkowski o Personalizzato.
- **Conteggi**. Misura chi-quadrato e Misura phi-quadrato.
- **Binaria**. Distanza euclidea, Distanza euclidea al quadrato, 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 prossimità (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

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 correlate** richiede che vengano trattate come variabili continue, in modo che le correlazioni (valori uguali per casi differenti) siano risolte in modo ottimale.

Condizionalità. Consente di specificare quali comparazioni sono significative. 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 dell'oggetto negativi**, se appropriata per i dati disponibili.

Scaling multidimensionale: Opzioni

È possibile impostare le opzioni per l'analisi scaling multidimensionale:

Visualizza. 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.

Funzioni 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.

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

Capitolo 32. Statistiche dei rapporti

La procedura Statistica dei rapporti offre un elenco completo di statistiche di riepilogo 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 sulla statistica dei rapporti dall'output e salvare i risultati in un file esterno.

Esempio. Esiste un buon grado di uniformità nel rapporto tra prezzo di valutazione 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.

Considerazioni sui dati della statistica dei rapporti

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

Ipotesi. Le variabili che definiscono il numeratore e il denominatore del rapporto dovrebbero essere variabili di scala con valori positivi.

Per ottenere la statistica dei rapporti

1. Dai menu, scegliere:
   Analizza > **Statistiche descrittive** > **Rapporto...**
2. Selezionare una variabile numeratore.
3. 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 Visualizzatore.
- 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 dei rapporti

Tendenza centrale. Le misure della tendenza centrale sono le 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 della divisione del risultato per il numero totale di rapporti.
- **Media pesata.** Il risultato della divisione della media del numeratore per la media del denominatore. La media pesata è anche la media dei rapporti pesati dal denominatore.

- **Intervalli di confidenza.** Vengono visualizzati gli intervalli di confidenza per la media, la mediana e la media pesata (se necessario). Specificare un valore maggiore o uguale a 0 e minore di 100 come intervallo di confidenza.

Dispersione. Queste statistiche misurano la quantità di variazione, o diffusione, 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 centrato sulla mediana.** (Coefficiente di variazione) È il risultato dell'espressione delle radici quadrate medie della deviazione dalla mediana come una percentuale della mediana.
- **COV centrato sulla 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 sottrazione del rapporto minimo dal rapporto massimo.
- **Minimo.** Il minimo è il rapporto più piccolo.
- **Massimo.** Il massimo è il rapporto più grande.

Indice di concentrazione. Il coefficiente di concentrazione misura la percentuale dei rapporti che rientrano 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**. L'estremità inferiore dell'intervallo è uguale a $(1 - 0.01 \times \text{valore}) \times \text{mediana}$ e l'estremità superiore è uguale a $(1 + 0.01 \times \text{valore}) \times \text{mediana}$.

Capitolo 33. Curve ROC

Questa procedura è utile per valutare la prestazione 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.

Considerazioni sui dati della curva ROC

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 uno stimatore di 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*.

Ipotesi. Si suppone che i numeri crescenti sulla scala dello stimatore 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

1. Dai menu, scegliere:
   Analizza > **Curva ROC...**
2. Selezionare una o più variabili di probabilità per il test.
3. Selezionare una variabile di stato.
4. Individuare il valore *positivo* per la variabile di stato.

Curva ROC: Opzioni

È possibile selezionare le seguenti opzioni per l'analisi ROC:

Classificazione . Consente di specificare se il valore di interruzione 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 mancati. Consente di specificare le modalità di gestione dei valori mancanti.

Capitolo 34. Simulazione

I modelli predittivi, ad esempio la regressione lineare, richiedono un insieme di input noti per prevedere un risultato o un valore di destinazione. In molte applicazioni del mondo reale, tuttavia, i valori di input sono incerti. La simulazione consente di tenere conto dell'incertezza negli input per i modelli predittivi e valutare la possibilità di vari risultati del modello in presenza di tale incertezza. Ad esempio, si ha un modello di profitto che include il costo dei materiali come un input, ma vi è incertezza in tale costo a causa della volatilità del mercato. È possibile utilizzare la simulazione per modellare tale incertezza e determinare l'effetto che essa ha sul profitto.

La simulazione in IBM SPSS Statistics utilizza il metodo Monte Carlo. Gli input incerti vengono modellati con le distribuzioni della probabilità (ad esempio, la distribuzione triangolare) e i valori simulati per tali input vengono generati a partire da tali distribuzioni. Gli input i cui valori sono noti, vengono tenuti fissi sui valori noti. Il modello predittivo viene valutato utilizzando un valore simulato per ogni input incerto e i valori fissi per gli input noti per calcolare la destinazione (o le destinazioni) del modello. Il processo viene ripetuto molte volte (in genere, decine di migliaia o centinaia di migliaia di volte), risultando in una distribuzione dei valori di destinazione che è possibile utilizzare per rispondere alle domande di natura probabilistica. Nel contesto di IBM SPSS Statistics, ogni ripetizione del processo genera un caso separato (record) di dati, composto dall'insieme di valori simulati per gli input incerti, dai valori degli input fissi e dalla destinazione o dalle destinazioni previste del modello.

È anche possibile simulare i dati in assenza di un modello predittivo specificando le distribuzioni di probabilità per le variabili che devono essere simulate. Ogni caso di dati generato consiste nella serie di valori simulati per le variabili specificate.

Per eseguire una simulazione, è necessario specificare i dettagli quali il modello predittivo, le distribuzioni della probabilità per gli input incerti, le correlazioni tra tali input e i valori per gli input fissi. Una volta specificati tutti i dettagli di una simulazione, è possibile eseguirla e, facoltativamente, salvare le specifiche in un file del **piano di simulazione**. È possibile condividere il piano di simulazione con altri utenti, che possono quindi eseguire la simulazione senza dover comprendere i dettagli del modo in cui è stata creata.

Sono disponibili due interfacce per utilizzare le simulazioni. Il Builder di simulazioni è un'interfaccia avanzata per gli utenti che progettano ed eseguono le simulazioni. Fornisce l'insieme completo di funzioni per la progettazione di una simulazione, per il salvataggio delle specifiche in un file del piano di simulazione, la specifica dell'output e l'esecuzione della simulazione. È possibile creare una simulazione in base a un file del modello IBM SPSS o a un insieme di equazioni personalizzate definite dall'utente nel Builder di simulazioni. Inoltre, è possibile caricare un piano di simulazione esistente nel Builder di simulazioni, modificare le impostazioni ed eseguire la simulazione, salvando facoltativamente il piano aggiornato. Per gli utenti che hanno un piano di simulazione e desiderano principalmente eseguire la simulazione, è disponibile un'interfaccia più semplice. Essa consente di modificare le impostazioni che consentono di eseguire la simulazione in condizioni differenti, ma non offre le funzioni complete del Builder di simulazioni per la progettazione delle simulazioni.

Progettazione di una simulazione in base a un file del modello

1. Dai menu, scegliere:

   Analizza > **Simulazione...**
2. Fare clic su **Seleziona file di modello SPSS** e fare clic su **Continua**.
3. Aprire il file del modello.

   Il file del modello è un file XML contenente il modello PMML creato da IBM SPSS Statistics o IBM SPSS Modeler. Per ulteriori informazioni, consultare l'argomento “Scheda Modello” a pagina 180.

4. Nella scheda Simulazione (in Builder di simulazioni) specificare le distribuzioni di probabilità per gli input simulati e i valori per gli input fissi. Se il dataset attivo contiene i dati cronologici per gli input simulati, fare clic su **Adatta tutto** per determinare automaticamente la distribuzione che si adatta meglio ai dati per ciascuno di tali input e per determinare le correlazioni tra loro. Per ciascun input simulato che non viene adattato dai dati cronologici, è necessario specificare esplicitamente una distribuzione, selezionando un tipo di distribuzione e immettendo i parametri richiesti.
5. Fare clic su **Esegui** per eseguire la simulazione. Per impostazione predefinita, il piano di simulazione che specifica i dettagli della simulazione, viene salvato nell'ubicazione specificata nelle impostazioni di salvataggio.

Sono disponibili le seguenti opzioni:

- Modificare l'ubicazione del piano di simulazione salvato.
- Specificare correlazioni note tra gli input simulati.
- Calcolare automaticamente una tabella di contingenza di associazioni tra gli input categoriali e utilizzare queste associazioni quando i dati vengono generati per questi input.
- Specificare l'analisi di sensibilità per esaminare l'effetto che si ottiene modificando il valore di un input fisso o modificando un parametro di distribuzione di un input simulato.
- Specificare le opzioni avanzate, ad esempio l'impostazione del numero massimo di casi da generare o la richiesta del campionamento delle code.
- Personalizzare l'output.
- Salvare i dati simulati in un file di dati.

Progettazione di una simulazione in base a equazioni personalizzate

1. Dai menu, scegliere:

   Analizza > **Simulazione...**
2. Fare clic su **Digita le equazioni** e su **Continua**.
3. Fare clic su **Nuova equazione** nella scheda Modello (nel Builder di simulazioni) per definire ciascuna equazione nel modello predittivo.
4. Fare clic sulla scheda Simulazione e specificare le distribuzioni della probabilità per gli input simulati e i valori per gli input fissi. Se il dataset attivo contiene i dati cronologici per gli input simulati, fare clic su **Adatta tutto** per determinare automaticamente la distribuzione che si adatta meglio ai dati per ciascuno di tali input e per determinare le correlazioni tra loro. Per ciascun input simulato che non viene adattato dai dati cronologici, è necessario specificare esplicitamente una distribuzione, selezionando un tipo di distribuzione e immettendo i parametri richiesti.
5. Fare clic su **Esegui** per eseguire la simulazione. Per impostazione predefinita, il piano di simulazione che specifica i dettagli della simulazione, viene salvato nell'ubicazione specificata nelle impostazioni di salvataggio.

Sono disponibili le seguenti opzioni:

- Modificare l'ubicazione del piano di simulazione salvato.
- Specificare correlazioni note tra gli input simulati.
- Calcolare automaticamente una tabella di contingenza di associazioni tra gli input categoriali e utilizzare queste associazioni quando i dati vengono generati per questi input.
- Specificare l'analisi di sensibilità per esaminare l'effetto che si ottiene modificando il valore di un input fisso o modificando un parametro di distribuzione di un input simulato.
- Specificare le opzioni avanzate, ad esempio l'impostazione del numero massimo di casi da generare o la richiesta del campionamento delle code.
- Personalizzare l'output.
- Salvare i dati simulati in un file di dati.

Progettazione di una simulazione senza un modello predittivo

1. Dai menu, scegliere:

   Analizza > **Simulazione...**
2. Fare clic su **Crea dati simulati** e fare clic su **Continua**.
3. Nella scheda Modello (nel Builder di simulazioni), selezionare i campi che si desidera simulare. È possibile selezionare i campi dal dataset attivo oppure è possibile definire dei nuovi campi facendo clic su **Nuovo**.
4. Fare clic sulla scheda Simulazione e specificare le distribuzioni di probabilità per i campi che devono essere simulati. Se il dataset attivo contiene dei dati cronologici per uno qualsiasi di questi campi, fare clic su **Adatta tutto** per determinare automaticamente la distribuzione che si adatta meglio ai dati e per determinare le correlazioni tra i campi. Per i campi che non sono adatti ai dati cronologici, è necessario specificare esplicitamente una distribuzione selezionando un tipo di distribuzione e immettendo i parametri richiesti.
5. Fare clic su **Esegui** per eseguire la simulazione. Per impostazione predefinita, i dati simulati sono salvati nel nuovo dataset specificato nelle impostazioni di salvataggio. Inoltre, il piano di simulazione, che specifica i dettagli della simulazione, viene salvato nell'ubicazione specificata nelle impostazioni di salvataggio.

Sono disponibili le seguenti opzioni:

- Modificare l'ubicazione per i dati simulati oppure il piano di simulazione salvato.
- Specificare le correlazioni note tra i campi simulati.
- Calcolare automaticamente una tabella di contingenza di associazioni tra i campi categoriali e utilizzare queste associazioni quando i dati vengono generati per questi campi.
- Specificare l'analisi di sensibilità per esaminare l'effetto che si ottiene modificando un parametro di distribuzione per un campo simulato.
- Specificare le opzioni avanzate, ad esempio l'impostazione del numero di casi da generare.

Eseguire una simulazione da un piano di simulazione

Sono disponibili due opzioni per eseguire una simulazione da un piano di simulazione. È possibile utilizzare la finestra di dialogo Esegui simulazione, che è progettata principalmente per l'esecuzione da un piano di simulazione oppure è possibile utilizzare il Builder di simulazioni.

Per utilizzare la finestra di dialogo Esegui simulazione:

1. Dai menu, scegliere:

   Analizza > **Simulazione...**
2. Fare clic su **Apri un piano di simulazione esistente**.
3. Verificare che la casella di controllo **Apri in Builder di simulazioni** non sia selezionata e fare clic su **Continua**.
4. Aprire il piano di simulazione.
5. Fare clic su **Esegui** nella finestra di dialogo Esegui simulazione.

Per eseguire la simulazione dal Builder di simulazioni:

1. Dai menu, scegliere:

   Analizza > **Simulazione...**
2. Fare clic su **Apri un piano di simulazione esistente**.
3. Selezionare la casella di controllo **Apri in Builder di simulazioni** e fare clic su **Continua**.
4. Aprire il piano di simulazione.
5. Modificare le impostazioni desiderate nella scheda Simulazione.

6. Fare clic su **Esegui** per eseguire la simulazione.

Facoltativamente, è possibile eseguire quanto riportato di seguito.

- Impostare o modificare l'analisi di sensibilità per esaminare l'effetto che si ottiene modificando il valore di un input fisso o modificando un parametro di distribuzione di un input simulato.
- Riadattare le distribuzioni e le correlazioni per gli input simulati ai nuovi dati.
- Modificare la distribuzione per un input simulato.
- Personalizzare l'output.
- Salvare i dati simulati in un file di dati.

Builder di simulazioni

Il Builder di simulazioni offre l'insieme completo di funzioni per progettare ed eseguire le simulazioni. Consente di eseguire le attività generali riportate di seguito.

- Progettare ed eseguire una simulazione per un modello IBM SPSS definito in un file del modello PMML.
- Progettare ed eseguire una simulazione per un modello predittivo definito da un insieme di equazioni personalizzate specificate dall'utente.
- Progettare ed eseguire una simulazione che genera dati in assenza di un modello predittivo.
- Eseguire una simulazione in base a un piano di simulazione esistente, modificando facoltativamente le impostazioni del piano.

Scheda Modello

Per le simulazioni basate su un modello predittivo, la scheda Modello specifica l'origine del modello. Per le simulazioni che non includono un modello predittivo, la scheda Modello specifica i campi che devono essere simulati.

Seleziona file di modello SPSS. Questa opzione specifica che il modello predittivo è definito in un file del modello IBM SPSS. Un file del modello IBM SPSS è un file XML contenente il modello PMML creato da IBM SPSS Statistics o IBM SPSS Modeler. I modelli predittivi vengono creati dalle procedure, ad esempio Regressione lineare o Strutture ad albero delle decisioni all'interno di IBM SPSS Statistics, e possono essere esportati in un file del modello. È possibile utilizzare un file del modello differente facendo clic su **Sfoglia** e accedendo al file desiderato.

Modelli PMML supportati dalla simulazione

- Regressione lineare
- Modello lineare generalizzato
- Modello lineare generalizzato
- Regressione logistica binaria
- Regressione logistica multinomiale
- Regressione multinomiale ordinale
- Regressione di Cox
- Struttura ad albero
- Struttura ad albero boosted (C5)
- Discriminante
- Cluster TwoStep
- Cluster delle k medie
- Rete neurale
- Insieme di regole (elenco decisionale)

Nota:

- I modelli PMML che hanno più campi di destinazione (variabili) o suddivisioni non sono supportato per l'uso nella simulazione.
- I valori di input stringa in modelli di regressione logistica binaria sono limitati a 8 byte nel modello. Se si stanno inserendo input stringa nel dataset attivo, assicurarsi che i valori nei dati non superino 8 byte in lunghezza. I valori di dati che superano 8 byte sono esclusi dalla dist ribuzione categoriale associata per l'input e sono visualizzati come non corrispondenti nella tabell a di output Categorie senza corrispondenza.

Digita le equazioni per il modello. Questa opzione specifica che il modello predittivo è composto da una o più equazioni personalizzate che verranno create dall'utente. Creare le equazioni facendo clic su **Nuova equazione**. Viene visualizzata la finestra di dialogo Editor di equazioni. È possibile modificare le equazioni esistenti, copiarle per utilizzarle come modelli per le nuove equazioni, registrarle ed eliminarle.

- Il Builder di simulazioni non supporta i sistemi di equazioni simultanee o le equazioni che sono non lineari nella variabile di destinazione.
- Le equazioni personalizzate vengono valutate nell'ordine in cui vengono specificate. Se l'equazione per una determinata destinazione dipende da un'altra destinazione, l'altra destinazione deve essere definita da un'equazione precedente.

  Ad esempio, dato l'insieme di tre equazioni fornito di seguito, l'equazione per *profitto* dipende dai valori di *ricavo* e *spese*, quindi le equazioni per *ricavo* e *spese* devono precedere l'equazione per *profitto*.

```
ricavo = prezzo*volume
spese = fisso + volume*(costo_unitario_materiali + costo_unitario_lavoro)
profitto = ricavo - spese
```

Crea dati simulati senza un modello. Selezionare questa opzione per simulare i dati senza un modello predittivo. Specificare i campi che devono essere simulati selezionando i campi dal dataset attivo oppure facendo clic su **Nuovo** per definire i nuovi campi.

Editor di equazioni

L'Editor di equazioni consente di creare o di modificare un'equazione personalizzata per il modello predittivo.

- L'espressione per l'equazione può contenere dei campi dal dataset attivo o nuovi campi di input definiti dall'utente nell'Editor di equazioni.
- È possibile specificare le proprietà della destinazione, quali il livello di misurazione, le etichette dei valori e se l'output viene generato o meno per la destinazione.
- È possibile utilizzare le destinazioni delle equazioni definite in precedenza come input per l'equazione corrente, in modo da creare equazioni accoppiate.
- È possibile allegare un commento descrittivo all'equazione. I commenti vengono visualizzati insieme all'equazione nella scheda Modello.

1. Immettere il nome della destinazione. Se lo si desidera, fare clic su **Modifica** nella casella di testo Destinazione per aprire la finestra di dialogo Input definiti, che consente di modificare le proprietà predefinite della destinazione.
2. Per creare un'espressione, incollare i componenti nel campo Espressione numerica o digitarli direttamente nel campo Espressione numerica.

- È possibile creare un'espressione personalizzata utilizzando i campi del dataset attivo oppure è possibile definire nuovi input facendo clic sul pulsante **Nuovo**. Viene visualizzata la finestra di dialogo Input definiti.
- È possibile incollare le funzioni selezionando un gruppo nell'elenco Gruppo di funzioni e facendo doppio clic sulla funzione nell'elenco Funzioni (o selezionare la funzione e fare clic sulla freccia accanto all'elenco Gruppo di funzioni). Immettere i parametri indicati dai punti interrogativi. Il gruppo di funzioni etichettato **Tutti** offre un elenco di tutte le funzioni disponibili. Una breve descrizione della funzione attualmente selezionata viene visualizzata in un'area riservata della finestra di dialogo.

- Le costanti stringa devono essere incluse tra virgolette.
- Se i valori contengono numeri decimali, è necessario utilizzare un punto (.) come indicatore decimale.

Nota: la simulazione non supporta equazioni personalizzate con le destinazioni stringa.

Input definiti: Nella finestra di dialogo Input definiti è possibile definire nuovi input e impostare le proprietà delle destinazioni.

- Se un input da utilizzare in un'equazione non esiste nel dataset attivo, è necessario definirlo prima che possa essere utilizzato nell'equazione.
- Se si stanno simulando dei dati senza un modello predittivo, è necessario definire tutti gli input simulati che non esistono nel dataset attivo.

Nome. Specificare il nome per una destinazione o un input.

Destinazione. È possibile specificare il livello di misurazione di una destinazione. Il livello di misurazione predefinito è continuo. È anche possibile specificare se l'output verrà creato per questa destinazione. Ad esempio, per un insieme di equazioni accoppiate è possibile che si sia interessati solo all'output prodotto dalla destinazione per l'equazione finale, quindi si eliminerà l'output prodotto dalle altre destinazioni.

Input da simulare. Specifica che i valori dell'input verranno simulati in base a una distribuzione della probabilità specificata (la distribuzione della probabilità viene specificata nella scheda Simulazione). Il livello di misurazione determina l'insieme predefinito delle distribuzioni che vengono considerate durante la ricerca della distribuzione che meglio si adatta ai dati per l'input (facendo clic su **Adatta** o su **Adatta tutto** nella scheda Simulazione). Ad esempio, se il livello di misurazione è continuo, la distribuzione normale (appropriata per i dati continui) verrà considerata, ma la distribuzione binomiale non verrà considerata.

Nota: Selezionare un livello di misurazione di Stringa per gli input stringa. Gli input di stringa che devono essere simulati sono limitati alla distribuzione categoriale.

Input a valore fisso. Specifica che il valore dell'input è noto e verrà fissato sul valore noto. Gli input fissi possono essere di tipo numerico o stringa. Specificare un valore per l'input fisso. I valori stringa non devono essere racchiusi tra virgolette.

Etichette valori. È possibile specificare le etichette valori per le destinazioni, gli input simulati e gli input fissi. Le etichette valore vengono utilizzate nei grafici di output e nelle tabelle.

Scheda Simulazione

La scheda Simulazione specifica tutte le proprietà della simulazione diverse dal modello predittivo. Nella scheda Simulazione è possibile effettuare le attività generali riportate di seguito.

- Specificare le distribuzioni di probabilità per gli input simulati e i valori per gli input fissi.
- Specificare le correlazioni tra gli input simulati. Per gli input categoriali, è possibile specificare che le associazioni che esistono tra questi input nel dataset attivo vengano utilizzate quando i dati sono generati per questi input.
- Specificare le opzioni avanzate, ad esempio, un campionamento delle code e i criteri di adattamento delle distribuzioni ai dati cronologici.
- Personalizzare l'output.
- Specificare se salvare il piano di simulazione e, facoltativamente, salvare i dati simulati.

Campi simulati

Per eseguire una simulazione, ogni campi di input deve essere specificato come fisso o simulato. Gli input simulati sono quelli i cui valori sono incerti e verranno generati a partire da una distribuzione della probabilità specificata. Quando sono disponibili i dati cronologici per gli input da simulare, le

distribuzioni che si adattano meglio ai dati possono essere determinate automaticamente, insieme alle correlazioni tra tali input. Inoltre, è possibile specificare manualmente le distribuzioni o le correlazioni se i dati cronologici non sono disponibili o se sono necessarie delle distribuzioni o correlazioni specifiche.

Gli input fissi sono quelli i cui valori sono noti e rimangono costanti per ogni caso generato nella simulazione. Ad esempio, si dispone di un modello di regressione lineare per le vendite come una funzione di un numero di input tra cui il prezzo e si desidera tenere fisso il prezzo di mercato corrente. Quindi, si specificherà il prezzo come un input fisso.

Per le simulazioni basate su un modello predittivo, ogni predittore nel modello è un campo di input per la simulazione. Per le simulazioni che non includono un modello predittivo, i campi specificati nella scheda Modello sono gli input per la simulazione.

Adattamento automatico delle distribuzioni e calcolo delle correlazioni per gli input simulati. Se il dataset attivo contiene i dati cronologici per gli input che si desidera simulare, è possibile trovare automaticamente le distribuzioni che si adattano meglio ai dati per tali input e determinare le correlazioni tra loro. Le operazioni da eseguire vengono riportate di seguito.

1. Verificare che ciascuno degli input che si desidera simulare corrisponda al campo corretto nel dataset attivo. Gli input vengono elencati nella colonna Input e la colonna Adatta a visualizza il campo corrispondente nel dataset attivo. È possibile associare un input a un campo differente nel dataset attivo selezionando un elemento differente dall'elenco a discesa Adatta a.

   Il valore *-Nessuno-* nella colonna Adatta a indica che non è stato possibile associare automaticamente l'input a un campo nel dataset attivo. Per impostazione predefinita, gli input sono associati ai campi del dataset a livello di nome, misurazione e tipo (numerico o stringa). Se il dataset attivo non contiene i dati cronologici per l'input, specificare manualmente la distribuzione per l'input oppure specificare l'input come input fisso, come descritto di seguito.
2. Fare clic su **Adatta tutto**.

La distribuzione dell'adattamento più vicino e i parametri associati vengono visualizzati nella colonna Distribuzione, insieme a un grafico della distribuzione sovrapposto a un istogramma (o grafico grafico barre) dei dati cronologici. Le correlazioni tra gli input simulati vengono visualizzate nelle impostazioni di Correlazioni. È possibile esaminare i risultati dell'adattamento e personalizzare l'adattamento automatico della distribuzione per un particolare input selezionando la riga dell'input e facendo clic su **Dettagli adattamento**. Per ulteriori informazioni, consultare l'argomento “Dettagli adattamento” a pagina 185.

È possibile eseguire l'adattamento automatico della distribuzione per un particolare input selezionando la riga dell'input e facendo clic su **Adatta**. Le correlazioni per tutti gli input simulati che corrispondono ai campi nel dataset attivo vengono anch'esse calcolate automaticamente.

Nota: per gli input continui e ordinali, se non viene trovato un adattamento accettabile per una qualsiasi delle distribuzioni testate, la distribuzione empirica viene suggerita come adattamento più vicino. Per gli input continui, la distribuzione empirica è la funzione di distribuzione cumulativa dei dati cronologici. Per gli input ordinali, la distribuzione empirica è la distribuzione categoriale dei dati cronologici.

Specifica manuale delle distribuzioni. È possibile specificare manualmente la distribuzione della probabilità per gli input simulati selezionando la distribuzione dall'elenco a discesa **Tipo** e immettendo i parametri di distribuzione nella griglia Parametri. Dopo avere immesso i parametri di una distribuzione, un grafico di esempio della distribuzione, in base ai parametri specificati, verrà visualizzato adiacente alla griglia Parametri. Di seguito vengono riportate alcune note relative a distribuzioni particolari.

- **Categoriale.** La distribuzione categoriale descrive un campo di input che ha un numero fisso di valori, a cui si fa riferimento come categorie. A ogni categoria è associata una probabilità, in modo che la somma delle probabilità in tutte le categorie sia uguale a uno. Per immettere una categoria, fare clic sulla colonna a sinistra nella griglia Parametri e specificare la categoria come un valore numerico. Immettere la probabilità associata alla categoria nella colonna a destra.

Nota: Gli input categoriali da un modello PMML hanno le categorie che sono determinate dal modello e non possono essere modificate.

- **Binomiale negativa - Errori.** Descrive la distribuzione del numero di errori in una sequenza di prove prima che venga osservato un numero specificato di esiti positivi. Il parametro *thresh* è il numero specificato di esiti positivi e il parametro *prob* è la probabilità di esito positivo in una determinata prova.
- **Binomiale negativa - Prove.** Descrive la distribuzione del numero di prove necessarie prima che venga osservato un numero specificato di esiti positivi. Il parametro *thresh* è il numero specificato di esiti positivi e il parametro *prob* è la probabilità di esito positivo in una determinata prova.
- **Intervallo.** Questa distribuzione è composta da un insieme di intervalli, a ogni intervallo è assegnata una probabilità in modo che la somma delle probabilità in tutti gli intervalli sia uguale a 1. I valori in un determinato intervallo vengono tracciati da una distribuzione uniforme definita in tale intervallo. Gli intervalli sono specificati immettendo un valore minimo, un valore massimo e una probabilità associata.

  Ad esempio, si ritiene che il costo di una materia prima abbia il 40% di possibilità di rientrare nell'intervallo di $10 - $15 per unità e il 60% di possibilità di rientrare nell'intervallo di $15 - $20 per unità. L'utente modellerà il costo con una distribuzione dell'intervallo composta dai due intervalli [10 - 15] e [15 - 20], impostando la probabilità associata al primo intervallo su 0.4 e la probabilità associata al secondo intervallo su 0.6. Gli intervalli non devono essere continui e possono anche sovrapporsi. Ad esempio, è possibile specificare gli intervalli $10 - $15 e $20 - $25 o $10 - $15 e $13 - $16.
- **Weibull.** Il parametro *c* è un parametro di ubicazione facoltativo, che specifica dove è posizionata l'origine della distribuzione.

I parametri per le seguenti distribuzioni hanno un significato identico a quello nelle funzioni delle variabili random associate disponibili nella finestra di dialogo Calcola variabile: Bernoulli, beta, binomiale, esponenziale, gamma, lognormale, binomiale negativa (prove ed errori), normale, Poisson e uniforme. Per ulteriori informazioni, consultare l'argomento .

Specifica di input fissi. Specificare un input fisso selezionando Fisso dall'elenco a discesa **Tipo** nella colonna Distribuzione e immettendo un valore fisso. Il valore può essere di tipo numerico o stringa a seconda che l'input sia di tipo numerico o stringa. I valori stringa non devono essere racchiusi tra virgolette.

Specifica dei limini nei valori simulati. La maggior parte delle distribuzioni supportano la specifica di limiti superiori e inferiori nei valori simulati. È possibile specificare un limite inferiore immettendo un valore nella casella di testo **Min** ed è possibile specificare un limite inferiore immettendo un valore nella casella di testo **Max**.

Blocco degli input. Bloccando un input selezionando la casella di controllo nella colonna con l'icona di blocco, si esclude l'input dall'adattamento automatico della distribuzione. Ciò è particolarmente utile quando si specifica manualmente una distribuzione o un valore fisso e si desidera verificare che non sia influenzata dall'adattamento automatico della distribuzione. Il blocco è utile anche se si desidera condividere il piano di simulazione con gli utenti che lo eseguiranno nella finestra di dialogo Esegui simulazione e si desidera impedire che vengano modificati specifici input. A tal proposito, le specifiche per gli input bloccati non possono essere modificate nella finestra di dialogo Esegui simulazione.

Analisi di sensibilità. L'analisi di sensibilità consente di esaminare l'effetto delle modifiche sistematiche in un input fisso o in un parametro di distribuzione per un input simulato, generando un insieme dipendente di casi simulati - in effetti, una simulazione separata - per ogni valore specificato. Per specificare l'analisi di sensibilità, selezionare un input fisso o simulato e fare clic su **Analisi di sensibilità**. L'analisi di sensibilità è limitata a un solo input fisso o a un solo parametro di distribuzione per un input simulato. Per ulteriori informazioni, consultare l'argomento “Analisi di sensibilità” a pagina 186.

Icone dello stato di adattamento

Le icone nella colonna Adatta a, indicano lo stato di adattamento per ciascun campo di input.

Tabella 3. Icone di stato.

Icona	Descrizione
	Non è stata specificata alcuna distribuzione per l'input e l'input non è stato specificato come fisso. Per eseguire la simulazione, è necessario specificare una distribuzione per questo input o definirlo come fisso e specificare il valore fisso.
	L'input è stato adattato in precedenza a un campo che non esiste nel dataset attivo. Non è necessaria alcuna azione, a meno che non si desideri riadattare la distribuzione per l'input al dataset attivo.
	La distribuzione dell'adattamento più vicina è stata sostituita con una distribuzione alternativa dalla finestra di dialogo Dettagli adattamento.
	L'input viene impostato sulla distribuzione dell'adattamento più vicina.
	La distribuzione è stata specificata manualmente o le iterazioni dell'analisi di sensibilità sono state specificate per questo input.

Dettagli adattamento: Nella finestra di dialogo Dettagli adattamento vengono visualizzati i risultati dell'adattamento della distribuzione automatica per un particolare input. Le distribuzioni sono ordinate per bontà di adattamento, per cui la distribuzione dell'adattamento più vicino viene elencata per prima. È possibile sovrascrivere la distribuzione dell'adattamento più vicina selezionando il pulsante di scelta per la distribuzione desiderata nella colonna Utilizza. Selezionando un pulsante di scelta nella colonna Utilizza, viene visualizzato anche un grafico della distribuzione sovrapposto ad un istogramma (o grafico a barre) dei dati cronologici per l'input in questione.

Statistiche di adattamento. Per impostazione predefinita, per i campi continui il test di Anderson-Darling viene utilizzato per determinare la bontà dell'adattamento. In alternativa, solo per i campi continui, è possibile specificare il test di Kolmogorov-Smirnoff per la bontà dell'adattamento selezionando tale opzione nelle impostazioni di Opzioni avanzate. Per gli input continui, i risultati di entrambi i test vengono visualizzati nella colonna Statistiche di adattamento (A per Anderson-Darling e K per Kolmogorov-Smirnoff) e il test scelto viene utilizzato per ordinare le distribuzioni. Per gli input ordinali e nominali viene utilizzato il test del chi-quadrato. Inoltre, vengono visualizzati i valori p associati ai test.

Parametri. I parametri della distribuzione associati a ogni distribuzione adattata vengono visualizzati nella colonna Parametri. I parametri per le seguenti distribuzioni hanno un significato identico a quello nelle funzioni delle variabili random associate disponibili nella finestra di dialogo Calcola variabile: Bernoulli, beta, binomiale, esponenziale, gamma, lognormale, binomiale negativa (prove ed errori), normale, Poisson e uniforme. Per ulteriori informazioni, consultare l'argomento . Per la distribuzione categoriale, i nomi dei parametri sono le categorie e i valori dei parametri sono le probabilità associate.

Riadattamento con un insieme di distribuzioni personalizzate. Per impostazione predefinita, il livello di misurazione dell'input viene utilizzato per determinare l'insieme di distribuzioni considerate per l'adattamento della distribuzione automatica. Ad esempio, le distribuzioni continue, quali quelle lognormali e gamma, vengono considerate durante l'adattamento di un input continuo, ma le distribuzioni discrete, quali quelle Poisson e binomiali, non vengono considerate. È possibile scegliere un sottoinsieme delle distribuzioni predefinite selezionando le distribuzioni nella colonna Riadatta. Inoltre, è possibile sovrascrivere l'insieme predefinito di distribuzioni, selezionando un livello di misurazione differente nell'elenco a discesa **Tratta come (misura)** e scegliendo le distribuzioni nella colonna Riadatta. Fare clic su **Esegui riadattamento** per eseguire il riadattamento con l'insieme di distribuzioni personalizzate.

Analisi di sensibilità: L'analisi della sensibilità consente di esaminare l'effetto che si ottiene modificando un input fisso o un parametro di distribuzione per un input simulato in un insieme specificato di valori. Un insieme indipendente di casi simulati - in effetti, una simulazione separata - viene generato per ogni valore specificato, consentendo all'utente di esaminare l'effetto della modifica dell'input. Ogni insieme di casi simulati viene definito come un'**iterazione**.

Itera. Questa opzione consente di specificare l'insieme di valori sui quali verrà modificato l'input.

- Se si modifica il valore di un parametro di distribuzione, selezionare il parametro dall'elenco a discesa. Immettere l'insieme di valori nella griglia Valore parametro per iterazione. Facendo clic su **Continua**, i valori specificati verranno aggiunti alla griglia Parametri per l'input associato e un indice specifica il numero di iterazioni del valore.
- Per le distribuzioni Categoriale e Intervallo, è possibile modificare le probabilità rispettivamente delle categorie o degli intervalli, ma i valori delle categorie e degli endpoint degli intervalli non possono essere modificati. Selezionare una categoria o un intervallo dall'elenco a discesa e specificare l'insieme di probabilità nella griglia Valore parametro per iterazione. Le probabilità per le altre categorie o gli altri intervalli verranno adattate automaticamente.

Nessuna iterazione. Utilizzare questa opzione per annullare le iterazioni per un input. Facendo clic su **Continua**, le iterazioni verranno rimosse.

Correlazioni

I campi di input da simulare sono spesso notoriamente correlati, ad esempio altezza e peso. Le correlazioni tra gli input che verranno simulati devono essere prese in considerazione per essere certi che i valori simulati conservino tali correlazioni.

Ricalcola correlazioni durante l'adattamento. Questa opzione specifica che le correlazioni tra gli input simulati vengono calcolate automaticamente durante l'adattamento delle distribuzione al dataset attivo mediante le azioni **Adatta tutto** o **Adatta** nelle impostazioni di Campi simulati.

Non ricalcolare correlazioni durante l'adattamento. Selezionare questa opzione se si desidera specificare manualmente le correlazioni e impedire che vengano sovrascritte durante l'adattamento automatico delle distribuzioni al dataset attivo. I valori immessi nella griglia Correlazioni devono essere compresi tra -1 e 1. Il valore 0 indica che non vi è alcuna correlazione tra la coppia associata di input.

Reimposta. Reimposta tutte le correlazioni su 0.

Utilizza tabella di contingenza a più vie adattata per gli input con una distribuzione categoriale. Per gli input con una distribuzione categoriale, è possibile calcolare automaticamente una tabella di contingenza a più vie dal dataset attivo che descrive le associazioni tra questi input. La tabella di contingenza viene quindi utilizzata quando i dati vengono generati per questi input. Se si sceglie di salvare il piano di simulazione, la tabella di contingenza viene salvata nel file del piano e viene utilizzata quando si esegue il piano.

- **Calcola tabella di contingenza dal dataset attivo.** Se si sta gestendo un piano di simulazione esistente che contiene una tabella di contingenza, è possibile ricalcolare la tabella di contingenza dal dataset attivo. Questa azione sovrascrive la tabella di contingenza dal file del piano caricato.
- **Utilizza tabella di contingenza dal piano di simulazione caricato.** Per impostazione predefinita, quando si carica un piano di simulazione che contiene una tabella di contingenza, viene utilizzata la tabella dal piano. È possibile ricalcolare la tabella di contingenza dal dataset attivo selezionando **Calcola tabella di contingenza dal dataset attivo**.

Opzioni avanzate

Numero massimo di casi. Specifica il numero massimo di casi dei dati simulati e i valori di destinazione associati da generare. Quando si specifica l'analisi di sensibilità, questo è il numero massimo di casi per ogni iterazione.

Destinazione per i criteri di arresto. Se il modello predittivo contiene più di una destinazione, è possibile selezionare la destinazione a cui si applicano i criteri di arresto.

Criteri di arresto. Queste opzioni specificano i criteri per arrestare la simulazione, potenzialmente prima che venga generato il numero massimo di casi consentiti.

- **Continua fino al numero massimo.** Specifica che i casi simulati verranno generati finché non viene raggiunto il numero massimo di casi.
- **Arresta quando le code sono campionate.** Utilizzare questa opzione quando si desidera verificare che il campionamento di una delle code di una distribuzione di destinazione specificata sia stato eseguito in modo adeguato. I casi simulati verranno generati finché non viene completato il campionamento delle code specificate o non viene raggiunto il numero massimo di casi. Se il modello predittivo contiene più destinazioni, selezionare la destinazione a cui verranno applicati questi criteri, dall'elenco a discesa **Destinazione per i criteri di arresto**.

  Tipo. È possibile definire il limite della regione della coda specificando un valore della destinazione, ad esempio 10.000.000 o un percentile, ad esempio, novantanovesimo percentile. Se si sceglie Valore nell'elenco a discesa **Tipo**, immettere il valore del limite nella casella di testo Valore e utilizzare l'elenco a discesa **Lato** per specificare se si tratta del limite della regione della coda destra o sinistra. Se si sceglie Percentile nell'elenco a discesa **Tipo**, immettere un valore nella casella di testo Percentile.

  Frequenza. Specificare il numero di valori della destinazione che devono trovarsi nella regione della coda per verificare che il campionamento della coda sia stato eseguito in modo adeguato. La creazione dei casi verrà arrestata quando viene raggiunto questo numero.
- **Arresta quando l'intervallo di confidenza della media rientra nella soglia specificata.** Utilizzare questa opzione quando si desidera verificare che la media di una determinata destinazione sia nota con un grado di precisione specificato. I casi simulati verranno generati finché non viene raggiunto il grado di precisione specificato o o non viene raggiunto il numero massimo di casi. Per utilizzare questa opzione, specificare un livello di confidenza e una soglia. I casi simulati verranno generati finché l'intervallo di confidenza associato al livello specificato è all'interno della soglia. Ad esempio, è possibile utilizzare questa opzione per specificare che i casi vengono generati finché l'intervallo di confidenza della media a livello di confidenza del 95% è entro il 5% del valore della media. Se il modello predittivo contiene più destinazioni, selezionare la destinazione a cui verranno applicati questi criteri, dall'elenco a discesa **Destinazione per i criteri di arresto**.

  Tipo di soglia. È possibile specificare la soglia come un valore numerico o una percentuale della media. Se si sceglie Valore nell'elenco a discesa **Tipo di soglia**, immettere la soglia nella casella di testo Soglia come valore. Se si sceglie Percentuale nell'elenco a discesa **Tipo di soglia**, immettere un valore nella casella di testo Soglia come percentuale.

Numero di casi da campionare. Specifica il numero di casi da utilizzare per l'adattamento automatico delle distribuzioni per gli input simulati nel dataset attivo. Se il dataset è molto grande, si consiglia di limitare il numero di casi utilizzati per l'adattamento della distribuzione. Se si seleziona **Limita a N casi**, verranno utilizzati i primi N casi.

Criteri della bontà di adattamento (input continui). Per gli input continui è possibile utilizzare il test di Anderson-Darling o il test di Kolmogorov-Smirnoff della bontà dell'adattamento per classificare le distribuzioni durante il loro adattamento per gli input simulati nel dataset attivo. Il test di Anderson-Darling è selezionato per impostazione predefinita ed è particolarmente consigliato quando si desidera ottenere il migliore adattamento possibile nelle regioni delle code.

Distribuzione empirica. Per gli input continui, la distribuzione empirica è la funzione di distribuzione cumulativa dei dati cronologici. È possibile specificare il numero di bin utilizzati per calcolare la distribuzione empirica degli input continui. Il valore predefinito è 100, il valore massimo è 1000.

Replica risultati. L'impostazione di un seme random consente di replicare la simulazione. Specificare un numero intero o fare clic su **Genera** per creare un numero intero pseudo-random compreso tra 1 e 2147483647, compresi. Il valore predefinito è 629111597.

Valori mancanti definiti dall'utente per gli input con una distribuzione categoriale. Questi controlli specificano se i valori mancanti definiti dall'utente di input con una distribuzione categoriale sono trattati come validi. I valori mancanti di sistema e i valori mancanti definiti dall'utente per tutti gli altri tipi di input sono sempre trattati come non validi. Tutti gli input devono avere dei valori validi per un caso per essere inclusi nell'adattamento della distribuzione, nel calcolo delle correlazioni e nel calcolo della tabella di contingenza facoltativa.

Funzioni di densità

Queste impostazioni consentono di personalizzare l'output per le funzioni di densità di probabilità e le funzioni di distribuzione cumulativa per le destinazioni continue, nonché i grafici a barre dei valori previsti per le destinazioni categoriali.

Funzione di densità di probabilità (PDF). La funzione di densità di probabilità visualizza la distribuzione dei valori di destinazione. Per le destinazioni continue consente di determinare la probabilità che la destinazione sia all'interno di una regione data. Per le destinazioni categoriali (destinazioni con un livello di misurazione nominale o ordinale), viene generato un grafico a barre che visualizza la percentuale di casi che rientrano in ciascuna categoria della destinazione. Altre opzioni per le destinazioni categoriali dei modelli PMML sono disponibili con l'impostazione Valori di categoria da segnalare descritta di seguito.

Per i modelli di cluster Two-Step e delle k medie, viene generato un grafico a barre dell'appartenenza al cluster.

Funzione di distribuzione cumulativa (CDF). La funzione di distribuzione cumulativa visualizza la probabilità che il valore della destinazione sia inferiore o uguale a un valore specificato. È disponibile solo per le destinazioni continue.

Posizioni dispositivo di scorrimento. È possibile specificare le posizioni iniziali delle linee di riferimento mobili su grafici PDF o CDF. I valori specificati per le linee inferiore e superiore fanno riferimento a posizioni lungo l'asse orizzontale, non a percentili. È possibile rimuovere la linea inferiore selezionando **-Infinità** oppure è possibile rimuovere la linea superiore selezionando **Infinità**. Per impostazione predefinita, le linee sono posizionate al 5° e al 95° percentile. Quando vengono visualizzate più funzioni di distribuzione su un singolo grafico (a causa di più destinazioni o risultati dalle iterazioni dell'analisi di sensibilità), il valore predefinito fa riferimento alla distribuzione per la prima iterazione o la prima destinazione.

Linee di riferimento (continue). È possibile richiedere diverse linee di riferimento verticali da aggiungere alle funzioni di densità di probabilità e alle funzioni di distribuzione cumulativa per le destinazioni continue.

- **Sigma.** È possibile aggiungere le linee di riferimento a più e meno di un numero specificato di deviazioni standard dalla media di una destinazione.
- **Percentili.** È possibile aggiungere delle linee di riferimento a uno o due valori di percentile della distribuzione di una destinazione, immettendo i valori nelle caselle di testo Basso e Alto. Ad esempio, il valore 95 nella casella di testo Alto rappresenta il 95° percentile, ovvero il valore al di sotto del quale ricade il 95% delle osservazioni. Allo stesso modo, il valore 5 nella casella di testo Basso rappresenta il 5° percentile, ovvero il valore al di sotto del quale ricade il 5% delle osservazioni.
- **Linee di riferimento personalizzate.** È possibile aggiungere le linee di riferimento nei valori specificati della destinazione.

Nota: Quando vengono visualizzate più funzioni di distribuzione su un singolo grafico (a causa di più destinazioni o risultati dalle iterazioni dell'analisi di sensibilità), le linee di riferimento vengono applicate solo alla distribuzione per la prima iterazione o per la prima destinazione. È possibile aggiungere delle linee di riferimento alle altre distribuzioni dalla finestra di dialogo Grafico: opzioni, a cui si accede dal grafico PDF o CDF.

Sovrapponi risultati da destinazioni continue separate. Nel caso di più destinazioni continue, specifica se le funzioni di distribuzione per tutte queste destinazioni vengono visualizzate in un solo grafico, con un grafico per le funzioni di densità di probabilità e un altro per le funzioni di distribuzione cumulativa. Quando questa opzione non è selezionata, i risultati di ciascuna destinazione verranno visualizzati in un grafico separato.

Valori di categoria da segnalare. Per i modelli PMML con destinazioni categoriali, il risultato del modello è un insieme di probabilità previste, una per ogni categoria, che il valore di destinazione rientri in ogni categoria. La categoria con la probabilità più alta viene considerata come la categoria prevista e viene utilizzata per la creazione del grafico a barre descritto per l'impostazione **Funzione di densità di probabilità** descritta in precedenza. Selezionando **Categoria prevista**, verrà generato il grafico a barre. Selezionando **Probabilità previste**, verranno generati gli istogrammi della distribuzione delle probabilità previste per ciascuna delle categorie della destinazione.

Raggruppamento per analisi di sensibilità. Le simulazioni che includono l'analisi di sensibilità generano un insieme indipendente di valori di destinazione previsti per ogni iterazione definita dall'analisi (un'iterazione per ogni valore dell'input che viene modificato). Quando sono presenti le iterazioni, il grafico a barre della categoria prevista per una destinazione categoriale viene visualizzato come un grafico a barre raggruppate che include i risultati di tutte le iterazioni. È possibile scegliere di raggruppare le categorie o le iterazioni.

Output

Grafici tornado. I grafici tornado sono dei grafici a barre che visualizzano le relazioni tra le destinazioni e gli input simulati utilizzando una varietà di metriche.

- **Correlazione della destinazione con l'input.** Questa opzione crea un grafico tornado dei coefficienti di correlazioni tra una destinazione data e ognuno dei suoi input simulati. Questo tipo di grafico tornado non supporta le destinazioni con un livello di misurazione nominale od ordinale oppure input simulati con una distribuzione categoriale.
- **Contributo alla varianza.** Questa opzione crea un grafico tornado che visualizza il contributo della varianza di una destinazione da ognuno dei suoi input simulati, consentendo all'utente di valutare in che misura ciascun input contribuisce all'incertezza globale nella destinazione. Questo tipo di grafico tornado non supporta le destinazioni con livelli di misurazione ordinali o nominali o input simulati con una qualsiasi delle seguenti distribuzioni: categoriale, Bernoulli, binomiale, Poisson o binomiale negativa.
- **Sensibilità della destinazione al cambiamento.** Questa opzione crea un grafico tornato che visualizza l'effetto sulla destinazione della modulazione di ogni input simulato per più o meno un numero specificato di deviazioni standard della distribuzione associata all'input. Questo tipo di grafico tornado non supporta le destinazioni con livelli di misurazione ordinali o nominali o input simulati con una qualsiasi delle seguenti distribuzioni: categoriale, Bernoulli, binomiale, Poisson o binomiale negativa.

Grafici a scatole delle distribuzioni destinazione. I grafici a scatole sono disponibili per le destinazioni continue. Selezionare **Sovrapponi risultati da destinazioni separate** se i modelli predittivi hanno più destinazioni continue e si desidera visualizzare i grafici a scatole per tutte le destinazioni in un solo grafico.

Grafici a dispersione delle destinazioni rispetto agli input. I grafici a dispersione delle destinazioni rispetto agli input simulati sono disponibili sia per le destinazioni continue che per quelle categoriali e includono i grafici a dispersione della destinazione con input continui e categoriali. Le dispersioni che interessano una destinazione categoriale o un input categoriale sono visualizzate come una mappa di calore.

Crea una tabella dei valori percentili. Per le destinazioni continue è possibile ottenere una tabella dei percentili specificati delle distribuzioni di destinazione. I quartili (il 25°, 50° e 75° percentile) suddividono le osservazioni in quattro gruppi di dimensioni uguali. Se si desidera avere un numero uguale di gruppi

diverso da quattro, selezionare **Intervalli** e specificare il numero. Selezionare **Percentili personalizzati** per specificare i singoli percentili, ad esempio il 99esimo percentile.

Statistiche descrittive delle distribuzioni destinazione. Questa opzione crea le tabelle di statistiche descrittive per le destinazioni continue e categoriali, nonché per gli input continui. Per le destinazioni continue la tabella include la media, la deviazione standard, la mediana, i valori minimo e massimo, l'intervallo di confidenza della media nel livello specificato e il quinto e 95esimo percentile della distribuzione di destinazione. Per le destinazioni categoriali la tabella include la percentuale di casi che rientrano in ciascuna categoria della destinazione. Per le destinazioni categoriali dei modelli PMML la tabella include anche la probabilità della media di ciascuna categoria della destinazione. Per gli input continui la tabella include la media, la deviazione standard e i valori minimo e massimo.

Correlazioni e tabella di contingenza per gli input. Questa opzione visualizza una tabella di coefficienti di correlazione tra gli input simulati. Quando gli input con le distribuzioni categoriali sono generati da una tabella di contingenza, viene visualizzata anche la tabella di contingenza dei dati generati per questi input.

Input simulati da includere nell'output. Per impostazione predefinita, tutti gli input simulati sono inclusi nell'output. È possibile escludere gli input simulati selezionati dall'output. In tal modo, verranno esclusi dai grafici tornado, dai grafici a dispersione e dall'output tabulare.

Intervalli di limite per le destinazioni continue. È possibile specificare l'intervallo di valori validi per una o più destinazioni continue. I valori fuori dall'intervallo specificato sono esclusi da tutto l'output e da tutte le analisi associate alle destinazioni. Per impostare un limite inferiore, selezionare **Inferiore** nella colonna Limite e immettere un valore nella colonna Minimo. Per impostare un limite superiore, selezionare **Superiore** nella colonna Limite e immettere un valore nella colonna Massimo. Per impostare sia un limite inferiore che un limite superiore, selezionare **Entrambi** nella colonna Limite e immettere i valori nelle colonne Minimo e Massimo.

Formati di visualizzazione. È possibile impostare il formato utilizzato durante la visualizzazione dei valori delle destinazioni e degli input (input fissi e input simulati).

Salvataggio di

Salva il file del piano per questa simulazione. È possibile salvare le specifiche correnti per la simulazione in un file del piano di simulazione. I file del piano di simulazione hanno estensione *.splan*. È possibile riaprire il piano nel Builder di simulazioni, apportare le modifiche, se lo si desidera, ed eseguire la simulazione. È possibile condividere il piano di simulazione con altri utenti, che possono quindi eseguirla nella finestra di dialogo Esegui simulazione. I piani di simulazione includono tutte le specifiche tranne le seguenti: impostazioni per Funzioni di densità; impostazioni di Output per i grafici e le tabelle; impostazioni di Opzioni avanzate per Adattamento, Distribuzione empirica e Seme random.

Salva i dati simulati come nuovo file di dati. È possibile salvare gli input simulati, gli input fissi e i valori di destinazione previsti in un file di dati SPSS Statistics, in un nuovo dataset nella sessione corrente o in un file Excel. Ogni caso (o riga) del file di dati è composto dai valori previsti delle destinazioni insieme agli input simulati e fissi che generano i valori di destinazione. Quando si specifica l'analisi di sensibilità, ogni iterazione dà luogo a un insieme continuo di casi che sono etichettati con il numero dell'iterazione.

Finestra di dialogo Esegui simulazione

La finestra di dialogo Esegui simulazione è progettata per gli utenti che hanno un piano di simulazione e desiderano principalmente eseguire una simulazione. Essa offre anche le funzioni necessarie per eseguire la simulazione in condizioni differenti. Consente di eseguire le attività generali riportate di seguito.

- Impostare o modificare l'analisi di sensibilità per esaminare l'effetto che si ottiene modificando il valore di un input fisso o modificando un parametro di distribuzione di un input simulato.

- Riadattare le distribuzioni della probabilità per gli input incerti (e le correlazioni tra tali input) ai nuovi dati.
- Modificare la distribuzione di un input simulato.
- Personalizzare l'output.
- Eseguire la simulazione.

Scheda Simulazione

La scheda Simulazione consente di specificare l'analisi di sensibilità, riadattare le distribuzioni di probabilità per gli input simulati e le correlazioni tra gli input simulati e i nuovi dati, nonché modificare la distribuzione della probabilità associata a un input simulato.

La griglia Input simulati contiene una voce per ogni campo di input definito nel piano di simulazione. Ogni voce visualizza il nome dell'input e il tipo di distribuzione della probabilità associato all'input, insieme a un grafico di esempio della curva di distribuzione associata. Inoltre, a ogni input è associata un'icona di stato (un cerchio colorato con un segno di spunta), che è utile quando si esegue il riadattamento delle distribuzioni ai nuovi dati. Inoltre, gli input possono includere un'icona di blocco, che indica che l'input è bloccato e non può essere modificato o riadattato ai nuovi dati nella finestra di dialogo Esegui simulazione. Per modificare un input bloccato, è necessario aprire il piano di simulazione nel Builder di simulazioni.

Ogni input è simulato o fisso. Gli input simulati sono quelli i cui valori sono incerti e verranno generati a partire da una distribuzione della probabilità specificata. Gli input fissi sono quelli i cui valori sono noti e rimangono costanti per ogni caso generato nella simulazione. Per utilizzare un determinato input, selezionare la voce relativa all'input nella griglia Input simulati.

Specifica dell'analisi di sensibilità

L'analisi di sensibilità consente di esaminare l'effetto delle modifiche sistematiche in un input fisso o in un parametro di distribuzione per un input simulato, generando un insieme dipendente di casi simulati - in effetti, una simulazione separata - per ogni valore specificato. Per specificare l'analisi di sensibilità, selezionare un input fisso o simulato e fare clic su **Analisi di sensibilità**. L'analisi di sensibilità è limitata a un solo input fisso o a un solo parametro di distribuzione per un input simulato. Per ulteriori informazioni, consultare l'argomento “Analisi di sensibilità” a pagina 186.

Riadattamento delle distribuzioni ai nuovi dati

Per riadattare automaticamente le distribuzioni della probabilità per gli input simulati (e le correlazioni tra gli input simulati) ai dati nel dataset attivo, effettuare quanto segue.

1. Verificare che ciascuno degli input del modello corrisponda al campo corretto nel dataset attivo. Ogni input simulato è adattato al campo nel dataset attivo specificato nell'elenco a discesa **Campo** associato all'input in questione. È possibile identificare facilmente gli input che non hanno corrispondenza, cercando gli input con un'icona di stato che include un segno di spunta con un punto interrogativo, come illustrato di seguito.

2. Modificare le corrispondenze dei campi necessarie selezionando **Adatta a un campo nel dataset**, quindi selezionando il campo nell'elenco.
3. Fare clic su **Adatta tutto**.

Per ogni input che è stato adattato, viene visualizzata la distribuzione che si adatta meglio ai dati, insieme a un grafico della distribuzione sovrapposto a un istogramma (o grafico a barre) dei dati

cronologici. Se non viene trovato un adattamento accettabile, viene utilizzata la distribuzione empirica. Per gli input che sono adattati alla distribuzione empirica, viene visualizzato solo un istogramma dei dati cronologici perché la distribuzione empirica è infatti rappresentata da tale istogramma.

Nota: per un elenco completo di icone di stato, consultare l'argomento “Campi simulati” a pagina 182.

Modifica delle distribuzioni della probabilità

È possibile modificare la distribuzione della probabilità per un input simulato e, facoltativamente, modificare un input simulato in un input fisso o viceversa.

1. Selezionare l'input, quindi selezionare **Imposta manualmente la distribuzione**.
2. Selezionare il tipo di distribuzione e specificare i parametri della distribuzione. Per modificare un input simulato in un input fisso, selezionare Fisso nell'elenco a discesa **Tipo**.

Dopo avere immesso i parametri di una distribuzione, il grafico di esempio della distribuzione (visualizzato nell'immissione dell'input) viene aggiornato per riflettere le modifiche. Per ulteriori informazioni su come specificare manualmente le distribuzioni della probabilità, consultare l'argomento “Campi simulati” a pagina 182.

Includi valori mancanti definiti dall'utente degli input categoriali in fase di adattamento. Specifica se i valori mancanti definiti dall'utente di input con una distribuzione categoriale sono trattati come validi quando si sta eseguendo il riadattamento ai nel dataset attivo. I valori mancanti di sistema e i valori mancanti definiti dall'utente per tutti gli altri tipi di input sono sempre trattati come non validi. Tutti gli input devono avere dei valori validi per un caso per essere inclusi nell'adattamento della distribuzione e nel calcolo delle correlazioni.

Scheda Output

La scheda Output consente di personalizzare l'output generato dalla simulazione.

Funzioni di densità. Le funzioni di densità sono lo strumento principale per sondare l'insieme di risultati ottenuti dalla simulazione.

- **Funzione di densità di probabilità.** La funzione di densità di probabilità visualizza la distribuzione dei valori di destinazione, consentendo di determinare la probabilità che la destinazione sia all'interno di una regione data. Per le destinazioni con un insieme fisso di risultati, ad esempio "servizio insufficiente", "servizio sufficiente", "servizio buono" e "servizio eccellente", viene generato un grafico a barre che visualizza la percentuale di casi che rientrano in ciascuna categoria della destinazione.
- **Funzione di distribuzione cumulativa.** La funzione di distribuzione cumulativa visualizza la probabilità che il valore della destinazione sia inferiore o uguale a un valore specificato.

Grafici tornado. I grafici tornado sono dei grafici a barre che visualizzano le relazioni tra le destinazioni e gli input simulati utilizzando una varietà di metriche.

- **Correlazione della destinazione con l'input.** Questa opzione crea un grafico tornado dei coefficienti di correlazioni tra una destinazione data e ognuno dei suoi input simulati.
- **Contributo alla varianza.** Questa opzione crea un grafico tornado che visualizza il contributo della varianza di una destinazione da ognuno dei suoi input simulati, consentendo all'utente di valutare in che misura ciascun input contribuisce all'incertezza globale nella destinazione.
- **Sensibilità della destinazione al cambiamento.** Questa opzione crea un grafico tornato che visualizza l'effetto sulla destinazione della modulazione di ogni input simulato per più o meno una deviazione standard della distribuzione associata all'input.

Grafici a dispersione delle destinazioni rispetto agli input. Questa opzione genera dei grafici a dispersione delle destinazioni rispetto agli input simulati.

Grafici a scatole delle distribuzioni destinazione. Questa opzione genera dei grafici a scatole delle distribuzioni di destinazione.

Tabella quartili. Questa opzione genera una tabella dei quartili delle distribuzioni di destinazione. I quartili di una distribuzione sono il 25esimo, il 50esimo e il 75esimo percentile della distribuzione e dividono le osservazioni in quattro gruppi della stessa dimensione.

Correlazioni e tabella di contingenza per gli input. Questa opzione visualizza una tabella di coefficienti di correlazione tra gli input simulati. Una tabella di contingenza di associazioni tra gli input con una distribuzione categoriale viene visualizzata quando il piano di simulazione specifica la generazione di dati categoriali da una tabella di contingenza.

Sovrapponi risultati da destinazioni separate. Se il modello predittivo che si sta simulando contiene più destinazioni, è possibile specificare se i risultati ottenuti da destinazioni separate vengono visualizzati in un solo grafico. Questa impostazione si applica alle funzioni di densità di probabilità, alle funzioni di distribuzione cumulativa e ai grafici a scatole. Ad esempio, se si seleziona questa opzione, le funzioni di densità di probabilità per tutte le destinazioni verranno visualizzate in un solo grafico.

Salva il file del piano per questa simulazione. È possibile salvare le modifiche apportate alla simulazione in un file del piano di simulazione. I file del piano di simulazione hanno estensione *.splan*. È possibile riaprire il piano nella finestra di dialogo Esegui simulazione o nel Builder di simulazioni. I piani di simulazione includono tutte le specifiche, tranne le impostazioni di output.

Salva i dati simulati come nuovo file di dati. È possibile salvare gli input simulati, gli input fissi e i valori di destinazione previsti in un file di dati SPSS Statistics, in un nuovo dataset nella sessione corrente o in un file Excel. Ogni caso (o riga) del file di dati è composto dai valori previsti delle destinazioni insieme agli input simulati e fissi che generano i valori di destinazione. Quando si specifica l'analisi di sensibilità, ogni iterazione dà luogo a un insieme continuo di casi che sono etichettati con il numero dell'iterazione.

Se è necessaria una maggiore personalizzazione dell'output di quella disponibile qui, si consiglia di eseguire la simulazione dal Builder di simulazioni. Per ulteriori informazioni, consultare l'argomento “Eseguire una simulazione da un piano di simulazione” a pagina 179.

Utilizzo dell'output del grafico dalla simulazione

Numerosi grafici generati da una simulazione hanno funzioni interattive che consentono di personalizzare la visualizzazione. Le funzioni interattive sono disponibili attivando (facendo doppio clic) l'oggetto grafico nel Visualizzatore di output. Tutti i grafici di simulazione sono visualizzazioni di tipo lavagna grafica.

Grafici Funzione di densità di probabilità per le destinazioni continue. Questo grafico ha due righe di riferimento verticali scorrevoli che dividono il grafico in regioni separate. La tabella sotto il grafico mostra la probabilità che la destinazione sia in ognuna delle regioni. Se più funzioni di densità vengono visualizzate nello stesso grafico, la tabella ha una riga separata per le probabilità associate a ciascuna funzione di densità. Ognuna delle righe di riferimento ha un dispositivo di scorrimento (triangolo invertito) che consente di spostare facilmente la linea. Sono disponibili numerose altre funzioni facendo clic sul pulsante **Grafici: opzioni** nel grafico. In particolare, è possibile impostare esplicitamente le posizioni dei dispositivi di scorrimento, aggiungere righe di riferimento fisse e modificare la vista del grafico da una curva continua in un istogramma o viceversa. Per ulteriori informazioni, consultare l'argomento “Grafico: opzioni” a pagina 194.

Grafici Funzione di distribuzione cumulativa per le destinazioni continue. Questo grafico ha le stesse due righe di riferimento verticali scorrevoli e la tabella associata descritte per il grafico Funzione di densità di probabilità illustrato in precedenza. Inoltre, consente di accedere alla finestra di dialogo Grafici: opzioni, in cui è possibile impostare esplicitamente le posizioni dei dispositivi di scorrimento, aggiungere

righe di riferimento fisse e specificare se la funzione di distribuzione cumulativa viene visualizzata come una funzione crescente (impostazione predefinita) o come una funzione decrescente. Per ulteriori informazioni, consultare l'argomento “Grafico: opzioni”.

Grafici a barre per le destinazioni categoriali con iterazioni dell'analisi di sensibilità. Per le destinazioni categoriali con iterazioni dell'analisi di sensibilità i risultati per la categoria di destinazione prevista vengono visualizzati come un grafico a barre raggruppate, che include i risultati di tutte le iterazioni. Il grafico include un elenco a discesa che consente di eseguire il raggruppamento sulla categoria o sull'iterazione. Per i modelli di cluster Two-Step e delle k medie, è possibile scegliere di eseguire il raggruppamento sul numero di cluster o sull'iterazione.

Grafici a scatole per più destinazioni con iterazioni dell'analisi di sensibilità. Per i modelli predittivi con più destinazioni continue e iterazioni dell'analisi di sensibilità, se si sceglie di visualizzare i grafici a scatole per tutte le destinazioni in un solo grafico, viene generato un grafico a scatole raggruppate. Il grafico include un elenco a discesa che consente di eseguire il raggruppamento sulla destinazione o sull'iterazione.

Grafico: opzioni

La finestra di dialogo Grafici: opzioni consente di personalizzare la visualizzazione dei grafici attivati delle funzioni di densità di probabilità e delle funzioni di distribuzione cumulativa generate da una simulazione.

Visualizza. L'elenco a discesa **Visualizza** si applica solo al grafico Funzione di densità di probabilità. Consente di alternare la visualizzazione del grafico da una curva continua in un istogramma. Questa funzione non è disponibile quando vengono visualizzate più funzioni di densità nello stesso grafico. In tal caso, le funzioni di densità possono solo essere visualizzate come curve continue.

Ordina. L'elenco a discesa **Ordina** si applica solo al grafico Funzione di distribuzione cumulativa. Specifica se la funzione di distribuzione cumulativa viene visualizzata come una funzione crescente (impostazione predefinita) o come una funzione decrescente. Quando viene visualizzata come una funzione decrescente, il valore della funzione in un determinato punto sull'asse orizzontale è la probabilità che la destinazione si trovi a destra di tale punto.

Posizioni dispositivo di scorrimento. È possibile impostare esplicitamente le posizioni delle righe di riferimento scorrevoli immettendo i valori nelle caselle di testo Superiore e Inferiore. È possibile rimuovere la linea a sinistra selezionando **-Infinito**, impostando in modo efficace la posizione su infinito negativo, ed è possibile rimuovere la linea a destra selezionando **Infinito**, impostando in modo efficace la sua posizione su infinito.

Righe di riferimento. È possibile aggiungere diverse linee di riferimento verticali fisse alle funzioni di densità di probabilità e alle funzioni di distribuzione cumulativa. Quando su un singolo grafico vengono visualizzate più funzioni (a causa di più destinazioni o risultati da iterazioni dell'analisi di sensibilità), è possibile specificare le specifiche funzioni a cui vengono applicate le linee.

- **Sigma.** È possibile aggiungere le linee di riferimento a più e meno di un numero specificato di deviazioni standard dalla media di una destinazione.
- **Percentili.** È possibile aggiungere delle linee di riferimento a uno o due valori di percentile della distribuzione di una destinazione, immettendo i valori nelle caselle di testo Basso e Alto. Ad esempio, il valore 95 nella casella di testo Alto rappresenta il 95° percentile, ovvero il valore al di sotto del quale ricade il 95% delle osservazioni. Allo stesso modo, il valore 5 nella casella di testo Basso rappresenta il 5° percentile, ovvero il valore al di sotto del quale ricade il 5% delle osservazioni.
- **Posizioni personalizzate.** È possibile aggiungere le righe di riferimento nei valori specificati lungo l'asse orizzontale.

Linee di riferimento etichetta. Questa opzione controlla se le etichette sono applicate alle linee di riferimento selezionate.

Le linee di riferimento vengono rimosse deselezionando l'opzione associata nella finestra di dialogo Grafici: opzioni e facendo clic su **Continua**.

Avvisi

Queste informazioni sono state sviluppate per prodotti e servizi offerti negli Stati Uniti.

IBM può non offrire i prodotti, i servizi o le funzioni presentati in questo documento in altri paesi. Consultare il proprio rappresentante locale IBM per informazioni sui prodotti ed i servizi attualmente disponibili nella propria zona. Qualsiasi riferimento ad un prodotto, programma o servizio IBM non implica o intende dichiarare che solo quel prodotto, programma o servizio IBM può essere utilizzato. Qualsiasi prodotto funzionalmente equivalente al prodotto, programma o servizio che non violi alcun diritto di proprietà intellettuale IBM può essere utilizzato. È comunque responsabilità dell'utente valutare e verificare la possibilità di utilizzare altri prodotti, programmi o servizi non IBM.

IBM può avere applicazioni di brevetti o brevetti in corso relativi all'argomento descritto in questo documento. La fornitura del presente documento non concede alcuna licenza a tali brevetti. È possibile inviare per iscritto richieste di licenze a:

Director of Commercial Relations
IBM Europe
Schoenaicher
D 7030 Boeblingen
Deutschland

Per richieste di licenze relative ad informazioni double-byte (DBCS), contattare il Dipartimento di Proprietà Intellettuale IBM nel proprio paese o inviare richieste per iscritto a:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

Il seguente paragrafo non si applica al Regno Unito o a qualunque altro paese in cui tali dichiarazioni sono incompatibili con le norme locali: IBM (INTERNATIONAL BUSINESS MACHINES CORPORATION) FORNISCE LA PRESENTE PUBBLICAZIONE "NELLO STATO IN CUI SI TROVA" SENZA GARANZIE DI ALCUN TIPO, ESPRESSE O IMPLICITE, IVI INCLUSE, A TITOLO DI ESEMPIO, GARANZIE IMPLICITE DI NON VIOLAZIONE, DI COMMERCIABILITÀ E DI IDONEITÀ PER UNO SCOPO PARTICOLARE. Alcuni stati non consentono la rinuncia ad alcune garanzie espresse o implicite in determinate transazioni, pertanto, la presente dichiarazione può non essere applicabile.

Queste informazioni potrebbero contenere imprecisioni tecniche o errori tipografici. Le modifiche alle presenti informazioni vengono effettuate periodicamente; tali modifiche saranno incorporate nelle nuove pubblicazioni della pubblicazione. IBM si riserva il diritto di apportare miglioramenti e/o modifiche al prodotto o al programma descritto nel manuale in qualsiasi momento e senza preavviso.

I riferimenti in queste informazioni a siti Web non IBM vengono forniti solo per comodità e non implicano in alcun modo l'approvazione di tali siti Web. I materiali presenti in tali siti Web non sono parte dei materiali per questo prodotto IBM e l'utilizzo di tali siti Web è a proprio rischio.

IBM può utilizzare o distribuire qualsiasi informazione fornita in qualsiasi modo ritenga appropriato senza incorrere in alcun obbligo verso l'utente.

Coloro che detengano la licenza su questo programma e desiderano avere informazioni su di esso allo scopo di consentire: (i) uno scambio di informazioni tra programmi indipendenti ed altri (compreso questo) e (ii) l'utilizzo reciproco di tali informazioni, dovrebbe rivolgersi a:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Tali informazioni possono essere disponibili, in base ad appropriate clausole e condizioni, includendo in alcuni casi, il pagamento di una tassa.

Il programma concesso in licenza descritto nel presente documento e tutto il materiale concesso in licenza disponibile sono forniti da IBM in base alle clausole dell'Accordo per Clienti IBM (IBM Customer Agreement), dell'IBM IPLA (IBM International Program License Agreement) o qualsiasi altro accordo equivalente tra le parti.

Qualsiasi dato sulle prestazioni qui contenuto è stato determinato in un ambiente controllato. Pertanto, i risultati ottenuti in altri ambienti operativi possono notevolmente variare. Alcune misurazioni possono essere state effettuate su sistemi del livello di sviluppo e non vi è alcuna garanzia che tali misurazioni resteranno invariate sui sistemi generalmente disponibili. Inoltre, alcune misurazioni possono essere state stimate tramite estrapolazione. I risultati reali possono variare. Gli utenti del presente documento dovranno verificare i dati applicabili per i propri ambienti specifici.

Le informazioni relative a prodotti non IBM sono ottenute dai fornitori di quei prodotti, dagli annunci pubblicati i da altre fonti disponibili al pubblico. IBM non ha testato quei prodotti e non può confermarne l'accuratezza della prestazione, la compatibilità o qualsiasi altro reclamo relativo ai prodotti non IBM. Le domande sulle funzionalità dei prodotti non IBM devono essere indirizzate ai fornitori di tali prodotti.

Tutte le dichiarazioni relative all'orientamento o alle intenzioni future di IBM sono soggette a modifica o a ritiro senza preavviso e rappresentano solo mete e obiettivi.

Questa pubblicazione contiene esempi di dati e prospetti utilizzati quotidianamente nelle operazioni aziendali. Pertanto, per maggiore completezza, gli esempi includono nomi di persone, società, marchi e prodotti. Tutti i nomi contenuti nel manuale sono fittizi e ogni riferimento a nomi e indirizzi reali è puramente casuale.

Ogni copia o qualsiasi parte di questi programmi di esempio o qualsiasi lavoro derivato, devono contenere le seguenti informazioni relative alle leggi sul diritto d'autore:

Questa pubblicazione contiene esempi di dati e prospetti utilizzati quotidianamente nelle operazioni aziendali. Pertanto, per maggiore completezza, gli esempi includono nomi di persone, società, marchi e prodotti. Tutti i nomi contenuti nel manuale sono fittizi e ogni riferimento a nomi e indirizzi reali è puramente casuale.

Ogni copia o qualsiasi parte di questi programmi di esempio o qualsiasi lavoro derivato, devono contenere le seguenti informazioni relative alle leggi sul diritto d'autore:

© (nome della società) (anno). Parti di questo codice derivano dai Programmi di Esempio di IBM Corp.

Marchi

IBM, il logo IBM e ibm.com sono marchi di International Business Machines Corp., registrati in molte giurisdizioni del mondo. Altri nomi di prodotti e servizi possono essere marchi di IBM o di altre società. Un elenco corrente dei marchi IBM è disponibile sul web in “Copyright and trademark information” all'indirizzo 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.

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 in altri paesi.

UNIX è un marchio della The Open Group negli Stati Uniti e/o negli altri paesi.

Java e tutti i marchi e i logo basati su Java sono marchi o marchi registrati di Oracle e/o delle sue associate.

Indice analitico

C

D

E

F

G

H

I

K

L

M

N

P

Q

R

S

T

U

V

W

Z

Stampato in Italia