IBM SPSS Statistics Base 24

Comunicado

Antes de usar estas informações e o produto suportado por elas, leia as informações nos “Avisos” na página 209.

Informações sobre o produto

Esta edição aplica-se à versão 24, liberação 0, modificação 0 do IBM SPSS Statistics e a todas as liberações e modificações subsequentes até que seja indicado de outra forma em novas edições.

Índice

Capítulo 1. Livro de códigos 1
Guia Saída do Codebook 1
Guia Estatísticas do Codebook 3

Capítulo 2. Frequências 5
Estatísticas de Frequências 5
Gráficos de Frequências. 7
Formato de Frequências 7

Capítulo 3. Descritivos 9
Opções Descritivas 9
Recursos Adicionais do Comando DESCRIPTIVES 10

Capítulo 4. Explorar 11
Explorar Estatísticas 12
Explorar Gráficos 12
Explorar Transformações de Potência 12
Opções de Exploração 13
Recursos Adicionais do Comando EXAMINE . . . 13

Capítulo 5. Crosstabs 15
Camadas de Crisstabs 16
Gráfico de barras agrupadas de Crosstabs 16
Crosstabs exibindo variáveis de camada em camadas da tabela 16
Estatísticas de Crosstabs 16
Exibição da célula de Crosstabs. 18
Formato de tabela de Crosstabs. 19

Capítulo 6. Resumir 21
Opções de Sumarização 21
Estatísticas de Sumarização 22

Capítulo 7. Médias 25
Opções de Médias 25

Capítulo 8. Cubos OLAP 29
Estatísticas de Cubos OLAP 29
Diferenças de Cubos OLAP 31
Título de Cubos OLAP 32

Capítulo 9. Teste t 33
Testes t 33
Teste-T de amostras independentes 33
Definir Grupos para Teste-T de Amostras Independentes 34
Opções de Teste T de Amostras Independentes 34
Teste-T de amostras em pares 34
Opções de Teste T de Amostras Emparelhadas. . 35
Recursos adicionais do comando T-TEST . . . 35
Teste-T de uma amostra 35
Opções de Teste T de Uma Amostra 36
Recursos adicionais do comando T-TEST . . . 36
Recursos adicionais do comando T-TEST 36

Capítulo 10. One-Way ANOVA 39
Contrastes One-Way ANOVA 39
Testes Post Hoc do One-Way ANOVA 40
Opções de One-Way ANOVA 41
Recursos Adicionais do Comando ONEWAY . . . 42

Capítulo 11. Análise Univariate GLM . . 43
Modelo de GLM. 44
Termos de compilação 45
Soma dos Quadrados 45
Contrastes GLM 46
Tipos de Contraste 46
Gráficos de perfil GLM 47
Opções de GLM 47
Recursos Adicionais do Comando UNIANOVA 48
Comparações Posteriori do GLM 48
Opções de GLM 50
Recursos Adicionais do Comando UNIANOVA 50
Salvamento de GLM 51
Opções de GLM 52
Recursos Adicionais do Comando UNIANOVA . . 52

Capítulo 12. Correlações Bivariadas . . 55
Opções de Correlações Bivariadas 55
Recursos Adicionais dos Comandos CORRELATIONS e NONPAR CORR 56

Capítulo 13. Correlações parciais . . . 57
Opções de Correlações Parciais 57
Recursos Adicionais do Comando PARTIAL CORR 58

Capítulo 14. Distâncias 59
Medidas de Dissimilaridade de Distâncias 59
Medidas de Similaridade de Distâncias 60
Recursos Adicionais do Comando PROXIMITIES . . 60

Capítulo 15. Modelos lineares 61
Para obter um modelo linear 61
Objetivos 61
Básico 62
Seleção de Modelo 63
Combinações 63
Avançado 64
Opções de Modelo 64
Sumarização do Modelo 64
Preparação de Dado Automático 64
Importância do Preditor 65
Predito por Observado 65
Resíduos 65
Valores Discrepantes 66
Efeitos 66
Coeficientes 66

Médias Estimadas 67
Sumarização de Construção de Modelo 67

Capítulo 16. Regressão linear 69
Métodos de Seleção de Variável de Regressão Linear 70
Regra do Conjunto de Regressão Linear 70
Gráficos de Regressão Linear 71
Regressão Linear: Salvando novas variáveis . . . 71
Estatísticas de Regressão Linear 73
Opções de regressão linear 73
Recursos Adicionais do Comando REGRESSION . . 74

Capítulo 17. Regressão ordinal 75
Opções de Regressão Ordinal 76
Saída de Regressão Ordinal 76
Modelo de Localização de Regressão Ordinal . . . 77
Termos de compilação 77
Modelo de Escala de Regressão Ordinal 78
Termos de compilação 78
Recursos Adicionais do Comando PLUM 78

Capítulo 18. Curva de Estimação . . . 79
Modelos de Curva de Estimação 80
Salvar na Curva de Estimação 80

Capítulo 19. Regressão por quadrados mínimos parciais 83
Modelo 84
Opções 85

Capítulo 20. Análise do vizinho mais próximo 87
Vizinhos 89
Recursos 90
Partições 90
Salvar 91
Saída 91
Opções 92
Visualização do Modelo 92
Espaço da Variável 92
Importância da Variável 94
Peers 94
Distâncias de Vizinho Mais Próximo 94
Mapa de Quadrante 94
Log de erro de seleção de variável. 95
Log de erro de seleção k 95
Log de Erro de k e Seleção de variável 95
Tabela de Classificação 95
Sumarização de Erro 95

Capítulo 21. Análise discriminante . . . 97
Análise Discriminante para Definir Intervalo . . . 98
Selecionar Casos para Análise Discriminante . . . 98
Estatística de Análise Discriminante 98
Método Stepwise para Análise Discriminante . . . 99
Classificação para Análise Discriminante 99
Salvamento de Análise Discriminante 100
Recursos Adicionais do Comando DISCRIMINANT 100

Capítulo 22. Análise fatorial 103
Selecionar Casos para Análise Fatorial 104
Descritivas de Análise Fatorial. 104
Extração de Análise Fatorial 104
Rotação de Análise Fatorial. 105
Escores de Análise Fatorial 106
Opções de Análise Fatorial 106
Recursos Adicionais do Comando FACTOR . . . 106

Capítulo 23. Escolhendo um Procedimento para Clusterização. . . 107

Capítulo 24. Análise de cluster de duas etapas 109
Opções de Análise de Cluster TwoStep 110
Saída da Análise de Cluster TwoStep 111
O Visualizador de Cluster 112
Visualizador de Cluster 112
Navegando no Visualizador de Cluster 115
Filtrando Registros 116

Capítulo 25. Análise de clusters hierárquica 119
Método de Análise de Cluster Hierárquica. . . . 119
Estatísticas de Análise de Cluster Hierárquica . . 120
Gráficos de Análise de Cluster Hierárquica . . . 120
Análise de Cluster Hierárquica para Salvar Novas Variáveis 120
Recursos Adicionais da Sintaxe do Comando CLUSTER 120

Capítulo 26. Análise de cluster por K-médias 123
Eficiência de Análise de Cluster por K-Médias . . 124
Iteração de Análise de Cluster por K-Médias . . . 124
Salvar Análise de Cluster por K-Médias 124
Opções de Análise de Cluster por K-médias . . . 125
Recursos Adicionais do Comando QUICK CLUSTER 125

Capítulo 27. Testes Não Paramétricos 127
Testes Não paramétricos de uma amostra 127
Para Obter Testes Não Paramétricos de Uma Amostra 127
Guia Campos 127
Guia Configurações 128
Variáveis Adicionais de Comando de NPTESTS 130
Testes Não Paramétricos de Amostras Independentes 130
Para Obter Testes Não Paramétricos de Amostras Independentes 131
Guia Campos 131
Guia Configurações 131
Variáveis Adicionais de Comando de NPTESTS 132
Testes Não Paramétricos de Amostras Relacionadas 133
Para Obter Testes Não Paramétricos de Amostras Relacionadas 133
Guia Campos 133
Guia Configurações 134

Variáveis Adicionais de Comando de NPTESTS 135
Visualização de Modelo 136
Visualização do Modelo 136
Variáveis Adicionais de Comando de NPTESTS . . 140
Diálogos Anteriores 140
Teste qui-quadrado 141
Teste de Binômio 142
Teste de Execuções 144
Teste de Kolmogorov-Smirnov de uma amostra 145
Testes de duas amostras independentes. 146
Testes de Duas Amostras Relacionadas 147
Testes para diversas amostras independentes 148
Testes para várias amostras relacionadas . . . 150

Capítulo 28. Análise de Múltiplas Respostas. 153
Análise de Múltiplas Respostas 153
Definir Conjuntos de Múltiplas Respostas 153
Frequências de múltiplas respostas 154
Tabulações cruzadas de múltiplas respostas . . . 155
Intervalos de Definição de Crosstabs de Múltiplas Respostas 156
Opções de Crosstabs de Múltiplas Respostas 156
Recursos Adicionais do Comando MULT RESPONSE 157

Capítulo 29. Resultados do Relatório 159
Resultados do Relatório 159
Sumarizações de Relatórios em Linhas 159
Para Obter um Relatório Sumarização: Sumarizações em Linhas 159
Formato de Quebra de Coluna de Dados do Relatório 160
Linhas de Sumarização de Relatório para/Linhas de Sumarização Final 160
Opções de Quebra de Relatório 160
Opções de relatório 160
Layout de relatórios 161
Títulos de Relatório 161
Sumarizações de Relatórios em Colunas 161
Para Obter um Relatório Sumarização: Sumarizações em Colunas 162
Função de Sumarização de Colunas de Dados 162
Sumarização de Colunas de Dados para Total de Colunas 162
Formato da Coluna do Relatório 163
Sumarizações de Relatórios em Opções de Quebra de Colunas 163
Sumarizações de Relatórios em Opções de Colunas 163
Layout de Relatório para Sumarizações nas colunas 163
Recursos Adicionais do Comando REPORT . . . 164

Capítulo 30. Análise de confiabilidade 165
Estatística de Análise de Confiabilidade 166
Recursos Adicionais do Comando RELIABILITY 167

Capítulo 31. Escala multidimensional 169
Forma de Dados para Ajuste de Escala Multidimensional 170
Criação de Medida para Ajuste de Escala Multidimensional 170
Modelo de Ajuste de Escala Multidimensional . . 170
Opções de Ajuste de Escala Multidimensional . . 171
Recursos Adicionais do Comando ALSCAL . . . 171

Capítulo 32. Estatísticas de razão. . . 173
Estatísticas de razão 173

Capítulo 33. Curvas ROC 175
Opções de Curva ROC 175

Capítulo 34. Simulação 177
Para designar uma simulação baseada em um arquivo de modelo 177
Para designar uma simulação baseada em equações customizadas 178
Para designar uma simulação sem um modelo preditivo 179
Para executar uma simulação a partir de um plano de simulação 179
Construtor de Simulação 180
Guia Modelo 180
Guia Simulação. 182
Executar diálogo de simulação 191
Guia de Simulação 191
Guia de saída 192
Trabalhando com saída de gráfico a partir da simulação 193
Opções de Gráfico. 194

Capítulo 35. Modelagem Geoespacial 197
Selecionando Mapas 197
Selecionando um Mapa 198
Relacionamento Geoespacial 198
Configurar Sistema de Coordenadas. 198
Configurando a Projeção 199
Sistema de Projeção e de Coordenadas 199
Origens de Dados 199
Incluir uma Origem de Dados. 200
Associação de Dados e Mapa 200
Validar chaves 200
Regras de associação geoespacial 200
Definir Campos de Dados do Evento 201
Selecionar Campos 201
Saída 201
Salve 202
Construção de Regra 203
Categorização e Agregação 204
Predição temporal espacial 204
Selecionar Campos 204
Intervalos de Tempo 205
Agregação 206
Saída 206
Opções de Modelo 207
Salvar 207
Avançado 207

Conclusão 208

Avisos 209
Marcas comerciais 211

Índice Remissivo 213

Capítulo 1. Livro de códigos

Codebook relata as informações de dicionário - como nomes de variáveis, rótulos de variáveis, rótulos de valor, valores omissos - e estatísticas básicas para todas ou somente as variáveis e conjuntos de respostas múltiplas especificados no conjunto de dados ativo. Para variáveis nominais e ordinais e conjuntos de múltiplas respostas, as estatísticas básicas incluem contagens e porcentagens. Para variáveis de escala, estatísticas básicas incluem a média, o desvio padrão e quartis.

Nota: o Codebook ignora o status de arquivo dividido. Isto inclui grupos de arquivo dividido criados para imputação múltipla de valores omissos (disponíveis na opção de complemento de Valores Omissos).

Para Obter um Codebook

1. Nos menus, escolha:
   Analisar > **Relatórios** > **Codebook**
2. Clique na guia Variáveis.
3. Selecione uma ou mais variáveis e/ou conjuntos de múltiplas respostas.

Como opção, você pode:

- Controlar as informações de variável que são exibidas.
- Controlar as estatísticas que são exibidas (ou excluir todas as estatísticas básicas).
- Controlar a ordem na qual as variáveis e os conjuntos de múltiplas respostas são exibidos.
- Alterar o nível de medição para qualquer variável na lista de origem para alterar as estatísticas básicas exibidas. Consulte o tópico “Guia Estatísticas do Codebook” na página 3 para obter mais informações

Alterando o Nível de Medição

É possível alterar temporariamente o nível de medição para variáveis. (Não é possível alterar o nível de medição para conjuntos de múltiplas respostas. Eles são sempre tratados como nominais).

1. Clique com o botão direito em uma variável na lista de origem.
2. Selecione um nível de medição no menu pop-up.

Isso altera o nível de medição temporariamente. Em termos práticos, isso é útil apenas para variáveis numéricas. O nível de medição para variáveis de sequência de caracteres é restrito a nominal ou ordinal, que são tratadas da mesma forma pelo procedimento Codebook.

Guia Saída do Codebook

A guia Saída controla as informações de variáveis incluídas para cada variável e conjunto de múltiplas respostas, a ordem na qual as variáveis e conjuntos de múltiplas respostas são exibidos e o conteúdo da tabela de informações do arquivo opcional.

Informações da variável

Isso controla as informações do dicionário exibidas para cada variável.

Posição. Um número inteiro que representa a posição da variável na ordem do arquivo. Isso não está disponível para conjuntos de múltiplas respostas.

Rótulo. O rótulo descritivo associado à variável ou conjunto de múltiplas respostas.

Tipo. Tipo de dados fundamentais. Este é *Numérico*, *Sequência de Caracteres* ou *Conjunto de Múltiplas Respostas*.

Formato. O formato de exibição para a variável, como *A4*, *F8.2*ou *DATE11*. Isso não está disponível para conjuntos de múltiplas respostas.

Nível de medição. Os valores possíveis são *Nominal*, *Ordinal*, *Escala* e *Desconhecido*. O valor exibido é o nível de medição armazenado no dicionário e não é afetado por nenhuma substituição de nível de medição temporário especificado ao alterar o nível de medição na lista de variáveis de origem na guia Variáveis. Isso não está disponível para conjuntos de múltiplas respostas.

Nota: O nível de medição para variáveis numéricas pode ser "desconhecido" antes da primeira passagem de dados quando o nível de medição não tiver sido configurado explicitamente, como dados lidos a partir de uma origem externa ou variáveis recém-criadas. Consulte o tópico para obter mais informações.

Papel. Alguns diálogos suportam a capacidade de pré-selecionar variáveis para análise com base em papéis definidos.

Rótulo de valor. Rótulos descritivos associados a valores de dados específicos.

- Se Contagem ou Porcentagem for selecionado na guia Estatísticas, os rótulos de valor definidos serão incluídos na saída mesmo se você não selecionar rótulos de Valor aqui.
- Para conjuntos de múltiplas dicotomias, os "rótulos de valor" são os rótulos de variáveis para as variáveis elementares no conjunto ou os rótulos de valores contados, dependendo de como o conjunto estiver configurado. Veja o tópico para obter mais informações.

Valores omissos. Valores omissos definidos pelo usuário. Se Contagem ou Porcentagem for selecionado na guia Estatísticas, os rótulos de valor definidos serão incluídos na saída, mesmo se você não selecionar Valores omissos aqui. Isso não está disponível para conjuntos de múltiplas respostas.

Atributos customizados. Atributos variáveis customizados definidos pelo usuário. A saída inclui os nomes e os valores para quaisquer atributos variáveis customizados associados a cada variável. Consulte o tópico para obter mais informações. Isso não está disponível para conjuntos de múltiplas respostas.

Atributos reservados. Atributos de variáveis do sistema reservados. É possível exibir atributos de sistema, mas eles não deverão ser alterados. Os nomes de atributos do sistema começam com um sinal de dólar ($). Atributos de não exibição, com nomes que começam com um "@" ou "$@", não são incluídos. A saída inclui os nomes e os valores para quaisquer atributos do sistema associados a cada variável. Isso não está disponível para conjuntos de múltiplas respostas.

Informações do arquivo

A tabela de informações do arquivo opcional pode incluir qualquer um dos seguintes atributos de arquivo:

Nome do arquivo. Nome do arquivo de dados do IBM® SPSS Statistics. Se o conjunto de dados nunca tiver sido salvo no formato IBM SPSS Statistics, então não há nenhum nome do arquivo de dados. (Se não houver nenhum nome de arquivo exibido na barra de títulos da janela Editor de Dados, então o conjunto de dados ativo não possui um nome de arquivo).

Localização. Localização do diretório (pasta) do arquivo de dados IBM SPSS Statistics. Se o conjunto de dados nunca tiver sido salvo no formato IBM SPSS Statistics, então não há nenhuma localização.

Número de casos. Número de casos no conjunto de dados ativo. Este é o número total de casos, incluindo todos os casos que podem ser excluídos das estatísticas básicas devido às condições do filtro.

Rótulo. Esse é o rótulo de arquivo (se houver) definido pelo comando FILE LABEL.

Documentos. Texto do documento do arquivo de dados.

Status de ponderação. Se a ponderação estiver ativa, o nome da variável de ponderação será exibido. Veja o tópico para obter mais informações.

Atributos customizados. Atributos de arquivo de dados customizados definidos pelo usuário. Atributos de arquivo de dados definidos com o comando DATAFILE ATTRIBUTE.

Atributos reservados. Atributos de arquivo de dados do sistema reservados. É possível exibir atributos de sistema, mas eles não deverão ser alterados. Os nomes de atributos do sistema começam com um sinal de dólar ($). Atributos de não exibição, com nomes que começam com um "@" ou "$@", não são incluídos. A saída inclui os nomes e valores de quaisquer atributos de arquivo de dados do sistema.

Ordem de exibição da variável

As alternativas a seguir estão disponíveis para controlar a ordem na qual as variáveis e conjuntos de respostas múltiplas são exibidos.

Alfabética. Ordem alfabética por nome de variável.

Arquivo. A ordem na qual as variáveis aparecem no conjunto de dados (a ordem em que elas são exibidas no Editor de Dados). Na ordem ascendente, conjuntos de múltiplas respostas são exibidos por último, após todas as variáveis selecionadas.

Nível de medição. Ordenar por nível de medição. Isso cria quatro grupos de ordenação: nominais, ordinais, escala e desconhecido. Conjuntos de respostas múltiplas são tratados como nominais.

Nota: O nível de medição para variáveis numéricas pode ser "desconhecido" antes da primeira passagem de dados quando o nível de medição não tiver sido configurado explicitamente, como dados lidos a partir de uma origem externa ou variáveis recém-criadas.

Lista de variáveis. A ordem na qual as variáveis e os conjuntos de múltiplas respostas aparecem na lista de variáveis selecionadas na guia Variáveis.

Nome do atributo customizado. A lista de opções de ordenação também inclui os nomes de quaisquer atributos variáveis customizados definidos pelo usuário. Na ordem crescente, as variáveis que não tiverem o atributo são ordenadas para a parte superior, seguido por variáveis que possuem o atributo, mas nenhum valor definido para o atributo, seguido por variáveis com os valores definidos para o atributo em ordem alfabética dos valores.

Número máximo de categorias

Se a saída incluir rótulos de valor, contagens ou porcentagens para cada valor exclusivo, será possível suprimir essas informações da tabela se o número de valores exceder o valor especificado. Por padrão, essas informações são suprimidas se o número de valores exclusivos para a variável exceder 200.

Guia Estatísticas do Codebook

A guia Estatísticas permite controlar as estatísticas básicas que são incluídas na saída, ou suprimir a exibição de estatísticas básicas inteiramente.

Contagens e porcentagens

Para variáveis nominais e ordinais, conjuntos de múltiplas respostas e valores rotulados de variáveis de escala, as estatísticas disponíveis são:

Conta. A contagem ou o número de casos que possuem cada valor (ou intervalo de valores) de uma variável.

Porcentagem. A porcentagem de casos que possuem um valor específico.

Dispersão e tendência central

Para variáveis de escala, as estatísticas disponíveis são:

Média. Uma medida de tendência central. A média aritmética, a soma dividida pelo número de casos.

Desvio padrão. Uma medida de dispersão em torno da média. Em uma distribuição normal, 68% dos casos estão dentro de um desvio padrão da média e 95% dos casos estão dentro de dois desvios padrão. Por exemplo, se a idade média for 45, com um desvio padrão de 10, 95% dos casos estariam entre 25 e 65 em uma distribuição normal.

Quartis. Exibe os valores correspondentes aos 25⌂, 50⌂ e 75⌂ percentis.

Nota: É possível alterar temporariamente o nível de medição associado a uma variável (e, portanto, alterar as estatísticas básicas exibidas para essa variável) na lista de variáveis de origem na guia Variáveis.

Capítulo 2. Frequências

O procedimento Frequências fornece exibições estatísticas e gráficas que são úteis para descrever muitos tipos de variáveis. O procedimento Frequências é um bom lugar para começar a observar seus dados.

Para um relatório de frequência e um gráfico de barras, é possível organizar os valores distintos em ordem crescente ou decrescente, ou ordenar as categorias pelas suas frequências. O relatório de frequências pode ser suprimido quando uma variável possui muitos valores distintos. É possível rotular gráficos com frequências (o padrão) ou porcentagens.

Exemplo. Qual é a distribuição de clientes de uma empresa por tipo de mercado? Na saída, você pode ver que 37,5% de seus clientes estão em agências do governo, 24,9% estão em empresas, 28,1% estão em instituições acadêmicas e 9,4% estão no segmento de mercado de assistência médica. Para dados quantitativos contínuos, como receita de vendas, você pode ver que a venda média do produto é $3.576, com um desvio padrão de $1.078.

Estatísticas e gráficos. Contagens de frequência, percentagens e porcentagens cumulativas, média, mediana, modo, soma, desvio padrão, variância, intervalo, valores mínimo e máximo, erro padrão da média, assimetria e curtose (ambas com erros padrão), quartis, percentis especificados pelo usuário, gráficos de barras, gráficos de pizza e histogramas.

Considerações de Dados de Frequências

Dados. Utilize códigos ou sequências de caracteres numéricos para codificar variáveis categóricas (níveis de medição nominais ou ordinais).

Suposições. As tabulações e porcentagens fornecem uma descrição útil dos dados a partir de qualquer distribuição, especialmente para as variáveis com categorias ordenadas ou não ordenadas. A maioria das estatísticas básicas opcionais, como a média e o desvio padrão, baseia-se em teoria normal e é apropriada para variáveis quantitativas com distribuições simétricas. Estatísticas robustas, como a mediana, quartis e percentis, são apropriadas para variáveis quantitativas que podem ou não satisfazer a suposição de normalidade.

Para Obter Tabelas de Frequência

1. Nos menus, escolha:

   Analisar > **Estatísticas Descritivas** > **Frequências...**
2. Selecione um ou mais variáveis categóricas ou quantitativas.

Como opção, você pode:

- Clicar em **Estatísticas** para estatísticas descritivas para variáveis quantitativas.
- Clicar em **Gráficos** para gráficos de barras, gráficos de setores circulares e histogramas.
- Clicar em **Formatar** para a ordem na qual os resultados são exibidos.

Estatísticas de Frequências

Valores de Percentil. Valores de uma variável quantitativa que dividem os dados ordenados em grupos de modo que uma determinada porcentagem esteja acima e outra porcentagem esteja abaixo. Quartis (os 25⌂, 50⌂ e 75⌂ percentis) dividem as observações em quatro grupos de tamanhos iguais. Se desejar um número igual de grupos diferente de quatro, selecione **Pontos de corte para n grupos iguais.** Também é possível especificar percentis individuais (por exemplo, o 95⌂ percentil, o valor abaixo do qual 95% dos observações caem).

Tendência Central. As estatísticas que descrevem a localização da distribuição incluem a média, a mediana, o modo e a soma de todos os valores.

- *Média*. Uma medida de tendência central. A média aritmética, a soma dividida pelo número de casos.
- *Median*. O valor acima e abaixo do qual metade dos casos cai, o 50º percentil. Se houver um número par de casos, a mediana é a média dos dois casos intermediários quando estão classificados em ordem crescente ou decrescente. A mediana é uma medida da tendência central não sensível a valores excessivos (ao contrário da média, que pode ser afetada por alguns valores extremamente altos ou baixos).
- *Modo*. O valor que ocorre mais frequentemente. Se vários valores compartilharem a maior frequência da ocorrência, cada um deles será um modo. O procedimento Frequências relata apenas o menor desses diversos modos.
- *Sum*. A soma ou o total dos valores, em todos os casos com valores não omissos.

Dispersão. Estatísticas que medem a quantia de variação ou de dispersão nos dados incluem o desvio padrão, variância, intervalo, mínimo, máximo e o erro padrão da média.

- *Desvio Padrão*. Uma medida de dispersão em torno da média. Em uma distribuição normal, 68% dos casos estão dentro de um desvio padrão da média e 95% dos casos estão dentro de dois desvios padrão. Por exemplo, se a idade média for 45, com um desvio padrão de 10, 95% dos casos estariam entre 25 e 65 em uma distribuição normal.
- *Variância*. Uma medida de dispersão ao redor da média, igual à soma dos desvios quadrados da média dividido por um menor que o número de casos. A variância é medida em unidades que são o quadrado daquelas da própria variável.
- *Intervalo*. A diferença entre os valores maior e menor de uma variável numérica, o máximo menos o mínimo.
- *Mínimo*. O menor valor de uma variável numérica.
- *Máximo*. O maior valor de uma variável numérica.
- *Erro Padrão da Média*. Uma medida do quanto o valor da média pode variar de amostra para amostra obtida da mesma distribuição. Ela pode ser utilizada para comparar aproximadamente a média observada com um valor hipotético (ou seja, é possível concluir que os dois valores serão diferentes se a razão da diferença com o erro padrão for inferior a -2 ou superior a +2).

Distribuição. Assimetria e curtose são estatísticas que descrevem a forma e a simetria da distribuição. Essas estatísticas são exibidas com seus erros padrão.

- *Assimetria*. Uma medida da assimetria de uma distribuição. A distribuição normal é simétrica e tem um valor de assimetria de 0. Uma distribuição com assimetria positiva significativa tem um longo rodapé direito. Uma distribuição com assimetria negativa significativa tem um longo rodapé esquerdo. Como uma orientação, um valor de assimetria mais de duas vezes seu erro padrão é obtido para indicar uma partida de simetria.
- *Curtose*. Uma medida da extensão até a qual as observações se agrupam em torno de um ponto central. Para a uma distribuição normal, o valor da estatística da curtose é zero. A curtose positiva indica que, com relação a uma distribuição normal, as observações são agrupadas mais ao centro da distribuição e têm rastros mais finos até os valores extremos da distribuição, no ponto em que os rastros da distribuição leptocúrtica são mais espessos com relação a uma distribuição normal. Já a curtose negativa indica que, com relação a uma distribuição normal, as observações são menos agrupadas e possuem rodapés mais espessos até os valores extremos da distribuição, no ponto em que os rastros da distribuição platykurtic são mais finos com relação a uma distribuição normal.

Os valores são pontos médios do grupo. Se os valores em seus dados forem pontos médios de grupos (por exemplo, idades de todas as pessoas com mais de trinta são codificadas como 35), selecione essa opção para estimar a mediana e os percentis para os dados originais não agrupados.

Gráficos de Frequências

Tipo de Gráfico. Um gráfico de pizza exibe a contribuição de partes para um todo. Cada fatia de um gráfico de pizza corresponde a um grupo que é definido por uma variável de agrupamento única. Um gráfico de barras exibe a contagem para cada valor ou categoria distinta como uma barra separada, permitindo comparar categorias visualmente. Um histograma também possui barras, mas eles são representados ao longo de uma escala de intervalo igual. A altura de cada barra é a contagem de valores de uma variável quantitativa que caem no intervalo. Um histograma mostra a forma, o centro e a dispersão da distribuição. Uma curva normal sobreposta em um histograma ajuda a avaliar se os dados são normalmente distribuídos.

Valores de Gráfico. Para gráficos de barras, o eixo de escala pode ser rotulado por contagens ou porcentagens de frequência.

Formato de Frequências

Ordenar por. A tabela de frequências pode ser organizada de acordo com os valores reais nos dados ou de acordo com a contagem (frequência de ocorrência) desses valores, e pode ser organizada em uma ordem crescente ou decrescente. No entanto, se você solicitar um histograma ou percentis, Frequências supõe que a variável é quantitativa e exibe seus valores em ordem crescente.

Diversas Variáveis. Se você produzir tabelas de estatísticas para variáveis múltiplas, será possível exibir todas as variáveis em uma única tabela (**Comparar variáveis**) ou exibir uma tabela de estatísticas separada para cada variável (**Organizar saída por variáveis)**.

Suprimir tabelas com muitas categorias. Esta opção impede a exibição de tabelas com mais do que o número especificado de valores.

Capítulo 3. Descritivos

O procedimento Descritivos exibe estatísticas básicas univariadas de diversas variáveis em uma única tabela e calcula valores padronizados (escores *z*). As variáveis podem ser classificadas pelo tamanho de suas médias (em ordem crescente ou decrescente), em ordem alfabética ou pela ordem na qual você seleciona as variáveis (o padrão).

Quando escores *z* são salvos, eles são incluídos nos dados no Editor de Dados e estão disponíveis para gráficos, listagens de dados e análises. Quando as variáveis são gravadas em unidades diferentes (por exemplo, produto interno bruto per capita e em porcentagem), uma transformação de escore *z* colocará as variáveis em uma escala comum para facilitar a comparação visual.

Exemplo. Se cada caso em seus dados contiver os totais de vendas diários de cada membro da equipe de vendas (por exemplo, uma entrada para Bob, uma entrada para Kim e uma entrada para Brian) coletados diariamente durante vários meses, o procedimento Descritivos poderá calcular a média diária de vendas para cada membro da equipe e ordenar os resultados de vendas médias mais altas para vendas médias mais baixas.

Estatísticas. Tamanho da amostra, média, mínimo, máximo, desvio padrão, variância, intervalo, soma, erro padrão da média e curtose e assimetria com seus erros padrão.

Considerações de Dados Descritivos

Dados. Utilize variáveis numéricas após verificá-las graficamente para registrar erros, valores discrepantes anomalias de distribuição. O procedimento Descritivos é muito eficiente para arquivos grandes (milhares de casos).

Suposições. A maioria das estatísticas disponíveis (incluindo escores *z*) baseia-se em teoria normal e é apropriada para variáveis quantitativas (medidas de nível de intervalo ou de razão) com distribuições simétricas. Evite variáveis com categorias não ordenadas ou distribuições defasadas. A distribuição de escores *z* possui a mesma forma que a dos dados originais, portanto, calcular escores *z* não é uma correção para dados de problema.

Para Obter Estatísticas Descritivas

1. Nos menus, escolha:
   Analisar > **Estatísticas Descritivas** > **Descritivos...**
2. Selecione uma ou mais variáveis.

Como opção, você pode:

- Selecionar **Salvar valores padronizados como variáveis** para salvar escores *z* como novas variáveis.
- Clicar em **Opções** para estatísticas opcionais e a ordem de exibição.

Opções Descritivas

Média e Soma. A média, ou média aritmética, é exibida por padrão.

Dispersão. Estatísticas que medem a dispersão ou a variação nos dados incluem o desvio padrão, variância, intervalo, mínimo, máximo e o erro padrão da média.

- *Desvio Padrão*. Uma medida de dispersão em torno da média. Em uma distribuição normal, 68% dos casos estão dentro de um desvio padrão da média e 95% dos casos estão dentro de dois desvios padrão. Por exemplo, se a idade média for 45, com um desvio padrão de 10, 95% dos casos estariam entre 25 e 65 em uma distribuição normal.
- *Variância*. Uma medida de dispersão ao redor da média, igual à soma dos desvios quadrados da média dividido por um menor que o número de casos. A variância é medida em unidades que são o quadrado daquelas da própria variável.
- *Intervalo*. A diferença entre os valores maior e menor de uma variável numérica, o máximo menos o mínimo.
- *Mínimo*. O menor valor de uma variável numérica.
- *Máximo*. O maior valor de uma variável numérica.
- *média de elemento de suporte*. Uma medida do quanto o valor da média pode variar de amostra para amostra obtida da mesma distribuição. Ela pode ser utilizada para comparar aproximadamente a média observada com um valor hipotético (ou seja, é possível concluir que os dois valores serão diferentes se a razão da diferença com o erro padrão for inferior a -2 ou superior a +2).

Distribuição. Curtose e assimetria são estatísticas que caracterizam a forma e a simetria da distribuição. Essas estatísticas são exibidas com seus erros padrão.

- *Curtose*. Uma medida da extensão até a qual as observações se agrupam em torno de um ponto central. Para a uma distribuição normal, o valor da estatística da curtose é zero. A curtose positiva indica que, com relação a uma distribuição normal, as observações são agrupadas mais ao centro da distribuição e têm rastros mais finos até os valores extremos da distribuição, no ponto em que os rastros da distribuição leptocúrtica são mais espessos com relação a uma distribuição normal. Já a curtose negativa indica que, com relação a uma distribuição normal, as observações são menos agrupadas e possuem rodapés mais espessos até os valores extremos da distribuição, no ponto em que os rastros da distribuição platykurtic são mais finos com relação a uma distribuição normal.
- *Assimetria*. Uma medida da assimetria de uma distribuição. A distribuição normal é simétrica e tem um valor de assimetria de 0. Uma distribuição com assimetria positiva significativa tem um longo rodapé direito. Uma distribuição com assimetria negativa significativa tem um longo rodapé esquerdo. Como uma orientação, um valor de assimetria mais de duas vezes seu erro padrão é obtido para indicar uma partida de simetria.

Ordem de Exibição. Por padrão, as variáveis são exibidas na ordem em que elas forem selecionadas. Opcionalmente, é possível exibir as variáveis em ordem alfabética, por médias crescentes ou por médias decrescentes.

Recursos Adicionais do Comando DESCRIPTIVES

O idioma da sintaxe de comando também permite:

- Salvar escores padronizados (escores *z*) para algumas, mas não todas as variáveis (com o subcomando `VARIABLES`).
- Especificar nomes para novas variáveis que contêm os escores padronizados (com o subcomando `VARIABLES`).
- Excluir da análise casos com valores omissos para qualquer variável (com o subcomando `MISSING`).
- Ordenar as variáveis na exibição pelo valor de quaisquer estatísticas, não apenas pela média (com o subcomando `SORT`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 4. Explorar

O procedimento Explorar produz estatísticas básicas e exibições gráficas, tanto para todos os casos ou separadamente para grupos de casos. Há muito motivos para utilizar o procedimento Explorar - triagem de dados, identificação de valor discrepante, descrição, verificação de suposição e caracterização das diferenças entre as subpopulações (grupos de casos). A triagem de dados pode mostrar presença de valores incomuns, valores extremos e intervalos nos dados, ou outras peculiaridades. Explorar os dados pode ajudar a determinar se as técnicas de estatísticas que estiverem sendo consideradas para análise de dados são apropriadas. A exploração poderá indicar uma necessidade de transformar os dados se a técnica requerer uma distribuição normal. Ou você pode decidir necessidade de testes não paramétricos.

Exemplo. Observe a distribuição dos tempos de aprendizagem em labirinto dos ratos em quatro diferentes planejamentos de reforço de aprendizado. Para cada um dos quatro grupos, é possível ver se a distribuição de tempos é aproximadamente normal e se as quatro variâncias são iguais. Também é possível identificar os casos com os cinco maiores e os cinco menores tempos. Os boxplots e gráficos de ramos e folhas resumem graficamente a distribuição dos tempos de aprendizado para cada um dos grupos.

Estatísticas e gráficos. Média, mediana, 5% da média aparada, erro padrão, variância, desvio padrão, mínimo, máximo, intervalo, amplitude interquartil, assimetria e curtose e seus erros padrão, intervalo de confiança para a média (e o nível de confiança especificado), percentis, estimador M de Huber, estimador de onda de Andrews, estimador M redescendente de Hampel, estimador de biponderação de Tukey, os cinco maiores e os cinco menores valores, estatística de Kolmogorov-Smirnov com um nível de significância de Lilliefors para teste de normalidade, e a estatística de Shapiro-Wilk. Boxplots, gráficos de ramos e folhas, histogramas, gráficos de normalidade e gráficos de dispersão versus nível com testes e transformações de Levene.

Considerações de Explorar Dados

Dados. O procedimento Explorar pode ser utilizado para variáveis quantitativas (medidas de nível de intervalo ou razão). Uma variável de fator (utilizada para dividir os dados em grupos de casos) deve ter um número razoável de valores distintos (categorias). Estes valores podem ser sequência curta ou numéricos. A variável de rótulo case, utilizada para rotular valores discrepantes em boxplots, pode ser uma sequência curta, uma sequência longa (primeiros 15 bytes) ou numérica.

Suposições. A distribuição dos dados não precisa ser simétrica ou normal.

Para Explorar Seus Dados

1. Nos menus, escolha:

   Analisar > **Estatísticas Descritivas** > **Explorar...**
2. Selecione uma ou mais variáveis dependentes.

Como opção, você pode:

- Selecionar uma ou mais variáveis de fator, cujos valores definirão grupos de casos.
- Selecionar uma variável de identificação para rotular casos.
- Clicar em **Estatísticas** para estimadores robustos, valores discrepantes, percentis e tabelas de frequência.
- Clicar em **Gráficos** para histogramas, gráficos e testes de probabilidade normal e gráficos de dispersão versus nível com estatísticas de Levene.
- Clicar em **Opções** para o tratamento de valores omissos.

Explorar Estatísticas

Descritivos. Essas medidas de tendência central e de dispersão são exibidas por padrão. As medidas de tendência central indicam a localização da distribuição; elas incluem a média, a mediana e 5% da média aparada. As medidas de dispersão mostram a dissimilaridade dos valores; essas incluem erro padrão, variância, desvio padrão, mínimo, máximo, intervalo e a amplitude interquartil. As estatísticas descritivas também incluem medidas da forma da distribuição; assimetria e curtose são exibidas com seus erros padrão. O intervalo de confiança de nível de 95% para a média também é exibido; é possível especificar um nível de confiança diferente.

Estimadores M. Alternativas robustas para a média e mediana da amostra para estimar a localização. Os estimadores calculados diferem nas ponderações que eles aplicam aos casos. Estimador M de Huber, o estimador de onda de Andrews, o estimador-M redescendente de Hampel e o estimador de biponderação de Tukey.

Valores Discrepantes. Exibe os cinco maiores e os cinco menores valores com rótulos case.

Percentis. Exibe os valores para o 5º, 10º, 25º, 50º, 75º, 90º e 95º percentis.

Explorar Gráficos

Boxplots. Essas alternativas controlam a exibição de boxplots quando tiver mais de uma variável dependente. **Níveis do fator juntos** gera uma exibição separada para cada variável dependente. Em uma exibição, boxplots são mostrados para cada um dos grupos definidos por uma variável de fator. **Dependentes juntos** gera uma exibição separada para cada grupo definido por uma variável de fator. Em uma exibição, boxplots são mostrados lado a lado para cada uma das variáveis dependentes. Esta exibição é particularmente útil quando as diferentes variáveis representam uma única característica medida em momentos diferentes.

Descritivo. O grupo Descritivo permite escolher gráficos de ramos e folhas e histogramas.

Gráficos de normalidade com testes. Exibe gráficos de probabilidade normal e de probabilidade normal de tendência. A estatística de Kolmogorov-Smirnov, com um nível de significância de Lilliefors para teste de normalidade, é exibida. Se ponderações de não número inteiro forem especificadas, a estatística de Shapiro-Wilk será calculada quando o tamanho da amostra ponderada estiver entre 3 e 50. Para nenhuma ponderação ou ponderações de número inteiro, a estatística é calculada quando o tamanho da amostra ponderada estiver entre 3 e 5.000.

Dispersão vs. Nível com Teste de Levene. Controla a transformação de dados para gráficos de dispersão versus nível. Para todos os gráficos de dispersão versus nível, a inclinação da linha de regressão e os testes robustos de Levene para a homogeneidade da variância são exibidos. Se selecionar uma transformação, testes de Levene terão como base os dados transformados. Se nenhuma variável de fator for selecionada, os gráficos de dispersão versus nível não serão produzidos. **Estimação de potência** produz um gráfico dos logs naturais das amplitudes interquartis com relação aos logs naturais das medianas para todas as células, bem como uma estimativa da transformação de potência para alcançar variâncias iguais nas células. Um gráfico de dispersão versus nível ajuda a determinar a potência de uma transformação para estabilizar (tornar mais igual) variâncias entre os grupos. **Transformado** permite selecionar uma das alternativas de potência, talvez seguindo a recomendação da estimativa de potência, e produz gráficos de dados transformados. A amplitude interquartil e a mediana dos dados transformados são representados. **Não transformado** produz gráficos de dados brutos. Isso é equivalente a uma transformação com uma potência de 1.

Explorar Transformações de Potência

Estas são as transformações de potência para gráficos de dispersão versus nível. Para transformar dados, deve-se selecionar uma potência para a transformação. É possível escolher uma das seguintes alternativas:

- **Logarítmica natural.** Transformação de log natural. Esse é o padrão.
- **1/raiz quadrada.** Para cada valor de dados, o recíproco da raiz quadrada é calculado.
- **Recíproco.** O recíproco de cada valor de dados é calculado.
- **Raiz quadrada.** A raiz quadrada de cada valor de dados é calculada.
- **Quadrado.** Cada valor de dados é elevado ao quadrado.
- **Cubo.** Cada valor de dados é elevado ao cubo.

Opções de Exploração

Valores omissos. Controla o tratamento de valores omissos.

- **Excluir listwise dos casos.** Casos com valores omissos para qualquer variável dependente ou de fator são excluídos de todas as análises. Esse é o padrão.
- **Excluir casos entre pares.** Casos sem valores omissos para variáveis em um grupo (célula) serão incluídos na análise desse grupo. O caso pode ter valores omissos para variáveis utilizadas em outros grupos.
- **Valores de relatório.** Valores omissos para variáveis de fator são tratados como uma categoria separada. Toda a saída é produzida para esta categoria adicional. Tabelas de frequência incluem categorias de valores omissos. Valores omissos de uma variável de fator são incluídos, mas rotulados como omissos.

Recursos Adicionais do Comando EXAMINE

O procedimento Explorar utiliza a sintaxe de comando do `EXAMINE`. O idioma da sintaxe de comando também permite:

- Solicitar saída total e gráficos, além de saída e gráficos para grupos definidos pelas variáveis de fator (com o subcomando `TOTAL`).
- Especificar uma escala comum para um grupo de boxplots (com o subcomando `SCALE`).
- Especificar as interações de variáveis de fator (com o subcomando `VARIABLES`).
- Especificar percentis diferentes dos padrões (com o subcomando `PERCENTILES`).
- Calcular percentis de acordo com qualquer um dos cinco métodos (com o subcomando `PERCENTILES`).
- Especificar qualquer transformação de potência para gráficos de dispersão versus nível (com o subcomando `PLOT`).
- Especificar o número de valores extremos a serem exibidos (com o subcomando `STATISTICS`).
- Especificar parâmetros para os estimadores M e estimadores robustos de localização (com o subcomando `MESTIMATORS`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 5. Crosstabs

O procedimento Crosstabs forma tabelas de dois fatores e de diversos fatores e fornece uma variedade de testes e medidas de associação para tabelas de dois fatores. A estrutura da tabela e se as categorias são ordenadas determinam qual teste ou medida utilizar.

As estatísticas e medidas de associação de Crosstabs são calculadas somente para tabelas de dois fatores. Se você especificar uma linha, uma coluna e um fator de estrato (variável de controle), o procedimento de Crosstabs criará um painel de estatísticas e de medidas associadas para cada valor do fator de estrato (ou uma combinação de valores para duas ou mais variáveis de controle). Por exemplo, se *gender* for um fator de estrato para uma tabela de *married* (sim, não) com relação a *life* (se a vida é excitante, rotineira ou monótona), os resultados de uma tabela de dois fatores para mulheres serão calculados separadamente dos resultados dos homens e impressos como painéis um seguido do outro.

Exemplo. É provável que clientes de pequenas empresas sejam mais rentáveis, em termos de vendas de serviços (por exemplo, treinamento e consultoria), do que clientes de empresas maiores? Em uma tabulação cruzada, é possível ver que a maioria das pequenas empresas (com menos de 500 funcionários) gera altos lucros de serviço, ao passo que a maioria das grandes empresas (com mais de 2.500 funcionários) gera menos lucros de serviço.

Estatísticas e medidas de associação. Qui-quadrado de Pearson, qui-quadrado de razão de verossimilhança, teste de associação linear por linear, teste exato de Fisher, qui-quadrado corrigido de Yates, *r* de Pearson, rô de Spearman, coeficiente de contingência, fi, *V* de Cramér, lambdas simétrica e assimétrica, tau de Goodman e Kruskal, coeficiente de incerteza, gama, *d* de Somers, tau-*b* e tau-*c* de Kendall, coeficiente eta, kappa de Cohen, estimativa de risco relativa, razão de chances, teste de McNemar, estatísticas de Cochran e Mantel-Haenszel e estatísticas de proporções da coluna.

Considerações de Dados de Crosstabs

Dados. Para definir as categorias de cada variável de tabela, utilize valores de uma variável numérica ou de sequência de caracteres (oito ou menos bytes). Por exemplo, para *gender*, é possível codificar os dados como 1 e 2 ou como *male* e *female*.

Suposições. Algumas estatísticas e medidas assumem categorias ordenadas (dados ordinais) ou valores quantitativos (dados de intervalo ou de proporção), conforme discutido na seção sobre estatísticas. Outras são válidas quando as variáveis da tabela possuem categorias não ordenadas (dados nominais). Para as estatísticas baseadas em qui-quadrado (fi, *V* de Cramér e o coeficiente de contingência), os dados devem ser uma amostra aleatória de uma distribuição multinomial.

Nota: As variáveis ordinais podem ser códigos numéricos que representam categorias (por exemplo, 1 = *low*, 2 = *medium*, 3 = *high*) ou valores de sequência de caracteres. Entretanto, a ordem alfabética dos valores de sequência de caracteres é assumida para refletir a ordem real das categorias. Por exemplo, para uma variável de sequência de caracteres com os valores *low*, *medium*, *high*, a ordem das categorias é interpretada como *high*, *low*, *medium* - que não é a ordem correta. Em geral, é mais confiável utilizar códigos numéricos para representar dados ordinais.

Para Obter Tabulações Cruzadas

1. Nos menus, escolha:

   Analisar > **Estatísticas Descritivas** > **Crosstabs...**
2. Selecione uma ou mais variáveis de linha e de uma ou mais variáveis de coluna.

Como opção, você pode:

- Selecionar uma ou mais variáveis de controle.
- Clicar em **Estatísticas** para os testes e medidas de associação para tabelas ou subtabelas de dois fatores.
- Clicar em **Células** para obter valores observados e esperados, porcentagens e resíduos.
- Clicar em **Formatar** para controlar a ordem das categorias.

Camadas de Crisstabs

Se selecionar uma ou mais variáveis da camada, uma tabulação cruzada separada será produzida para cada categoria de cada variável de camada (variável de controle). Por exemplo, se tiver uma variável de linha, uma variável da coluna e uma variável de camada com duas categorias, você obterá uma tabela de dois fatores para cada categoria da variável de camada. Para criar outra camada de variáveis de controle, clique em **Avançar**. Subtabelas são produzidas para cada combinação de categorias para cada variável de primeira camada, para cada variável segunda camada, e assim por diante. Se estatísticas e medidas de associação forem solicitadas, elas se aplicarão a apenas subtabelas de dois fatores.

Gráfico de barras agrupadas de Crosstabs

Exibir gráficos de barras agrupadas. Um gráfico de barras agrupadas ajuda a sumarizar seus dados para grupos de casos. Há um cluster de barras para cada valor da variável que você especificou em Linhas. A variável que define as barras dentro de cada cluster é a variável que você especificou em Colunas. Há um conjunto de barras coloridas ou padronizadas de forma diferente para cada valor dessa variável. Se você especificar mais de uma variável em Colunas ou Linhas, um gráfico de barras agrupadas será produzido para cada combinação de duas variáveis.

Crosstabs exibindo variáveis de camada em camadas da tabela

Exibir variáveis de camada nas camadas da tabela. É possível optar por exibir as variáveis de camada (variáveis de controle) como camadas da tabela na tabela de tabulação cruzada. Isso permite criar visualizações que mostram as estatísticas gerais das variáveis de linha e da coluna, além de realizar drill down nas categorias das variáveis da camada.

Um exemplo que utiliza o arquivo de dados *demo.sav* (disponível no diretório de Amostras do diretório de instalação) é mostrado abaixo e foi obtido conforme a seguir:

1. Selecione *Categoria de renda em milhares (inccat)* como a variável de linha, *Possui PDA (ownpda)* como a variável de coluna e *Nível de Educação (ed)* como a variável de camada.
2. Selecione **Exibir variáveis de camada em camadas da tabela**.
3. Selecione **Coluna** no subdiálogo Exibição de Célula.
4. Execute o procedimento Crosstabs, clique duas vezes na tabela de tabulação cruzada e selecione **Ensino superior** na lista suspensa Nível de educação.

A visualização selecionada da tabela de tabulação cruzada mostra as estatísticas dos respondentes que tiverem ensino superior.

Estatísticas de Crosstabs

Qui-quadrado. Para tabelas com duas linhas e duas colunas, selecione **Qui-quadrado** para calcular o qui-quadrado de Pearson, o qui-quadrado de razão de verossimilhança, o teste exato de Fisher e o qui-quadrado corrigido de Yates (correção de continuidade). Para tabelas 2×2, o teste exato de Fisher é calculado quando uma tabela que é não resultante de linhas ou colunas omissas em uma tabela maior possui uma célula com uma frequência esperada menor que 5. O qui-quadrado corrigido de Yates é calculado para todas as outras tabelas 2×2. Para tabelas com qualquer número de linhas e colunas, selecione **Qui-quadrado** para calcular o qui-quadrado de Pearson e o qui-quadrado de razão de verossimilhança. Quando ambas as variáveis da tabela são quantitativas, **Qui-quadrado** produz o teste de associação linear por linear.

Correlações. Para tabelas nas quais as linhas e colunas contêm valores ordenados, **Correlações** gera o coeficiente de correlação e o rô de Spearman (somente dados numéricos). O rô de Spearman é uma medida da associação entre ordens de ranqueamento. Quando as variáveis de tabela (fatores) são quantitativas, **Correlações** gera o coeficiente de correlação de Pearson, *r*, que é uma medida de associação linear entre as variáveis.

Nominal. Para dados nominais (nenhuma ordem intrínseca, como católica, protestante e judaica), é possível selecionar **Coeficiente de contingência**, **Fi** (coeficiente) **e V de Cramér**, **Lambda** (lambdas simétricos e assimétricos e tau de Goodman e Kruskal, e **Coeficiente de incerteza**.

- *Coeficiente de contingência*. Uma medida de associação com base em qui-quadrado. O valor varia entre 0 e 1, com 0 indicando nenhuma associação entre as variáveis de linha e de coluna e valores próximos de 1 indicando um alto grau de associação entre as variáveis. O valor máximo possível depende do número de linhas e de colunas em uma tabela.
- *Fi e V de Cramer*. Fi é uma medida de associação baseada em qui-quadrado que envolve dividir a estatística qui-quadrado pelo tamanho da amostra e obter a raiz quadrada do resultado. V de Cramer é uma medida de associação baseada em qui-quadrado.
- *Lambda*. Uma medida de associação que reflete a redução proporcional de erro quando valores da variável independente são usados para prever valores da variável dependente. Um valor de 1 significa que a variável independente prediz perfeitamente a variável dependente. Um valor 0 significa que a variável independente não ajuda a prever a variável dependente.
- *Coeficiente de incerteza*. Uma medida de associação que indica a redução proporcional de erro quando os valores de uma variável são utilizados para prever valores da outra variável. Por exemplo, um valor de 0,83 indica que o conhecimento de uma variável reduz em 83% o erro ao prever valores da outra variável. O programa calcula ambas as versões simétrica e assimétrica do coeficiente de incerteza.

Ordinal. Para tabelas nas quais as linhas e colunas contêm valores ordenados, selecione **Gama** (ordem zero para tabelas de dois fatores e condicional para tabelas de 3 a 10 fatores), **Tau-b de Kendall** e **Tau-c de Kendall**. Para categorias de coluna de predição a partir de categorias de linha, selecione **d de Somers**.

- *Gamma*. Uma medida de associação simétrica entre duas variáveis ordinais que varia entre -1 e 1. Os valores próximos a um valor absoluto de 1 indicam um relacionamento forte entre duas variáveis. Valores próximos a 0 indicam pouco ou nenhum relacionamento. Para tabelas de duas vias, gamas de ordem zero são exibidas. Para tabelas de 3 a n vias, gamas condicionais são exibidos.
- *d de Somers*. Uma medida de associação entre duas variáveis ordinais que varia de -1 a 1. Os valores próximos de um valor absoluto de 1 indicam um forte relacionamento entre as duas variáveis e valores próximos de 0 indicam pouco ou nenhum relacionamento entre as variáveis. O d de Somers é uma extensão assimétrica de gama que difere apenas na inclusão do número de pares não relacionados à variável independente. Uma versão simétrica dessa estatística também é calculada.
- *Tau-b de Kendall*. Uma medida de correlação não paramétrica para variáveis ordinais ou classificadas que leva os empates em consideração. O sinal do coeficiente indica a direção do relacionamento e seu valor absoluto indica a intensidade, com os valores absolutos maiores indicando os relacionamentos mais fortes. Os valores possíveis variam de -1 a 1, mas um valor de -1 ou +1 pode ser obtido apenas a partir de tabelas quadradas.
- *Tau-c de Kendall*. Uma medida de associação não paramétrica para variáveis ordinais que ignora os empates. O sinal do coeficiente indica a direção do relacionamento e seu valor absoluto indica a intensidade, com os valores absolutos maiores indicando os relacionamentos mais fortes. Os valores possíveis variam de -1 a 1, mas um valor de -1 ou +1 pode ser obtido apenas a partir de tabelas quadradas.

Nominal por Intervalo. Quando uma variável é categórica e a outra é quantitativa, selecione **Eta**. A variável categórica deve ser codificada de forma numérica.

- *Eta*. Uma medida de associação que varia de 0 a 1, com 0 indicando nenhuma associação entre as variáveis de linha e de coluna e valores próximos de 1 indicando um alto grau de associação. O Eta é apropriado para uma variável dependente medida em uma escala de intervalo (por exemplo, rendimento) e para uma variável independente com um número limitado de categorias (por exemplo,

sexo). Dois valores de eta são calculados: um trata a variável de linha como a variável de intervalo, e o outro trata a variável de coluna como a variável de intervalo.

Kappa. O kappa de Cohen mede a concordância entre as avaliações de dois avaliadores quando ambos estiverem classificando o mesmo objeto. Um valor de 1 indica concordância perfeita. Um valor de 0 indica que a concordância não é melhor que a chance. O kappa é baseado em uma tabela quadrada na qual os valores da linha e da coluna representam a mesma escala. Qualquer célula que possuir valores observados para uma variável, mas não para outra, é designada a uma contagem de 0. O kappa não será calculado se o tipo de armazenamento de dados (sequência de caracteres ou numérico) não for o mesmo para as duas variáveis. Para a variável de sequência de caracteres, ambas as variáveis devem ter o mesmo comprimento definido.

Risco. Para tabelas 2 x 2, uma medida da intensidade da associação entre a presença de um fator e a ocorrência de um evento. Se o intervalo de confiança para a estatística incluir um valor de 1, não será possível supor que o fator está associado ao evento. A razão de chances pode ser utilizada como uma estimativa ou risco relativo quando a ocorrência do fator for rara.

McNemar. Um teste não paramétrico para duas variáveis dicotômicas relacionadas. Testa as mudanças nas respostas usando a distribuição qui-quadrado. Ele é útil para detectar mudanças nas respostas devido à intervenção experimental em designs de "antes e depois". Para tabelas quadradas maiores, o teste de simetria de McNemar-Bowker é relatado.

Estatísticas de Cochran e Mantel-Haenszel. As estatísticas de Cochran e Mantel-Haenszel podem ser utilizadas para testar a independência entre uma variável de fator dicotômico e uma variável de resposta dicotômica, condicional aos padrões de covariáveis definidos por uma ou mais variáveis de camada (controle). Observe que, enquanto outras estatísticas são calculadas camada por camada, as estatísticas de Cochran e Mantel-Haenszel são calculadas uma vez para todas as camadas.

Exibição da célula de Crosstabs

Para ajudar a descobrir padrões nos dados que contribuem com um teste qui-quadrado significativo, o procedimento Crosstabs exibe as frequências esperadas e três tipos de resíduos (desvios) que medem a diferença entre as frequências observadas e esperadas. Cada célula da tabela pode conter qualquer combinação de contagens, porcentagens e resíduos selecionados.

Contagens. O número de casos realmente observados e o número de casos esperados se as variáveis de linha e de coluna forem independentes umas das outras. É possível optar por ocultar contagens que forem menores que um número inteiro especificado. Valores ocultos serão exibidos como **<N**, em que **N** é o número inteiro especificado. O número inteiro especificado deve ser maior ou igual a 2, embora o valor 0 seja permitido e especifica que nenhuma contagem é ocultada.

Comparar proporções da coluna. Esta opção calcula as comparações entre pares de proporções da coluna e indica quais pares de colunas (para uma determinada linha) são significativamente diferentes. Diferenças significativas são indicadas na tabela de tabulação cruzada com formatação de estilo APA usando letras subscritas e são calculadas no nível de significância de 0,05. *Nota*: Se essa opção for especificada sem selecionar contagens observadas ou porcentagens de coluna, então as contagens observadas serão incluídas na tabela de tabulação cruzada, com as letras subscritas no estilo APA indicando os resultados dos testes de proporções da coluna.

- **Ajustar valores de p (método de Bonferroni).** Comparações entre pares de proporções da coluna utilizam a correção de Bonferroni, que ajusta o nível de significância observado para o fato de que diversas comparações são feitas.

Porcentagens. As porcentagens podem aumentar entre as linhas ou diminuir entre as colunas. As porcentagens do número total de casos representados na tabela (uma estrato) também estão disponíveis. *Nota*: Se **Ocultar contagens pequenas** for selecionada no grupo Contagens, então as porcentagens associadas a contagens ocultas também serão ocultadas.

Residuais. Resíduos brutos não padronizados fornecem a diferença entre os valores observados e esperados. Resíduos padronizados e padronizados ajustados também estão disponíveis.

- *Não padronizado*. A diferença entre um valor observado e o valor esperado. O valor esperado é o número de casos que você esperaria na célula, se não houvesse nenhum relacionamento entre as duas variáveis. Um resíduo positivo indica que há mais casos na célula do que teria se as variáveis de linha e de coluna fossem independentes.
- *Padronizado*. O resíduo dividido por uma estimativa do seu desvio padrão. Resíduos padronizados, também conhecidos como resíduos de Pearson, possuem uma média de 0 e um desvio padrão de 1.
- *Padronizado ajustado*. O resíduo de uma célula (valor observado menos esperado) dividido por uma estimativa do seu erro padrão. O resíduo padronizado resultante é expresso em unidades de desvio padrão acima ou abaixo da média.

Ponderações de Não Número Inteiro. As contagens de células são geralmente valores de números inteiros, já que eles representam o número de casos em cada célula. Mas se o arquivo de dados estiver atualmente ponderado por uma variável de ponderação com valores fracionários (por exemplo, 1,25), as contagens de células também poderão ser valores fracionários. É possível truncar ou arredondar antes ou depois de calcular as contagens de células ou utilizar contagens de células fracionárias para exibição de tabela e cálculos estatísticos.

- *Arredondar contagens de células*. As ponderações de caso são utilizadas no estado em que se encontram, e as ponderações acumuladas nas células são arredondadas antes de calcular quaisquer estatísticas.
- *Truncar contagens de células*. As ponderações de caso são utilizadas no estado em que se encontram, e as ponderações acumuladas nas células são truncadas antes de calcular quaisquer estatísticas.
- *Arredondar ponderações de caso*. As ponderações de caso são arredondadas antes do uso.
- *Truncar ponderações de caso*. As ponderações de caso são truncadas antes do uso.
- *Nenhum ajustamento*. As ponderações de caso são utilizadas no estado em que se encontram e contagens de célula fracionais são utilizadas. No entanto, quando Estatísticas Exatas (disponíveis apenas com a opção Testes Exatos) são solicitadas, as ponderações acumuladas nas células são truncadas ou arredondadas antes de calcular as estatísticas de teste exatas.

Formato de tabela de Crosstabs

É possível organizar as linhas em ordem crescente ou decrescente dos valores da variável de linha.

Capítulo 6. Resumir

O procedimento Sumarizar calcula estatísticas de subgrupo para variáveis dentro das categorias de uma ou mais variáveis de agrupamento. Todos os níveis da variável de agrupamento possuem tabulação cruzada efetuada. É possível escolher a ordem em que as estatísticas são exibidas. As estatísticas básicas de cada variável em todas as categorias também são exibidas. Os valores de dados em cada categoria podem ser listados ou suprimidos. Com conjuntos de dados grandes, é possível optar por listar apenas os primeiros *n* casos.

Exemplo. Qual é a quantia média de vendas de produtos por região e segmento de mercado do cliente? Você pode descobrir que a quantia média de vendas é ligeiramente mais elevada na região oeste do que em outras regiões, com clientes corporativos na região oeste produzindo uma quantia média mais alta de vendas.

Estatísticas. Soma, número de casos, média, mediana, mediana agrupada, erro padrão da média, mínimo, máximo, intervalo, valor da variável da primeira categoria da variável de agrupamento, valor da variável da última categoria da variável de agrupamento, desvio padrão, variância, curtose, erro padrão da curtose, assimetria, erro padrão da assimetria, porcentagem da soma total, porcentagem do total de *N*, porcentagem da soma em, porcentagem de *N* em, média geométrica e média harmônica.

Considerações de Dados de Sumarização

Dados. Variáveis de agrupamento são variáveis categóricas cujos valores podem ser numéricos ou de sequência de caracteres. O número de categorias deve ser razoavelmente pequeno. As outras variáveis devem ser capazes de serem ranqueadas.

Suposições. Algumas das estatísticas de subgrupo opcionais, como a média e o desvio padrão, baseiam-se em teoria normal e são apropriadas para variáveis quantitativas com distribuições simétricas. Estatísticas robustas, como a mediana e o intervalo, são apropriadas para variáveis quantitativas que podem ou não satisfazer a suposição de normalidade.

Para Obter Sumarizações de Caso

1. Nos menus, escolha:

   Analisar > **Relatórios** > **Sumarizações de Caso...**
2. Selecione uma ou mais variáveis.

Como opção, você pode:

- Selecionar uma ou mais variáveis de agrupamento para dividir os dados em subgrupos.
- Clicar em **Opções** para alterar o título de saída, incluir uma legenda abaixo da saída ou excluir casos com valores omissos.
- Clicar em **Estatísticas** para estatísticas opcionais.
- Selecionar **Exibir casos** para listar os casos em cada subgrupo. Por padrão, o sistema listará somente os primeiros 100 casos em seu arquivo. É possível aumentar ou diminuir o valor para **Limitar casos aos primeiros** *n* ou cancele a seleção desse item na lista todos os casos.

Opções de Sumarização

A sumarização permite alterar o título de sua saída ou incluir uma legenda que aparecerá abaixo da tabela de saída. É possível controlar a quebra de linha em títulos e legendas digitando `\n` onde desejar inserir uma quebra de linha no texto.

Também é possível optar por exibir ou suprimir subtítulos para totais e incluir ou excluir casos com valores omissos para qualquer uma das variáveis utilizadas em qualquer uma das análises. Normalmente é desejável denotar os casos omissos na saída com um ponto ou um asterisco. Insira um caractere, frase ou código que você deseja que apareça quando um valor estiver omisso; caso contrário, nenhum tratamento especial será aplicado aos casos omissos na saída.

Estatísticas de Sumarização

É possível escolher uma ou mais das seguintes estatísticas de subgrupo para as variáveis dentro de cada categoria de cada variável de agrupamento: soma, número de casos, média, mediana, mediana agrupada, erro padrão da média, mínimo, máximo, intervalo, valor da variável da primeira categoria da variável de agrupamento, valor da variável da última categoria da variável de agrupamento, desvio padrão, variância, curtose, erro padrão da curtose, assimetria, erro padrão da assimetria, porcentagem da soma total, porcentagem do total de *N*, porcentagem da soma em, porcentagem de *N* em, média geométrica e média harmônica. A ordem na qual as estatísticas aparecem na lista de Estatísticas de Célula é a ordem na qual elas serão exibidas na saída. As estatísticas de sumarização também são exibidas para cada variável em todas as categorias.

First. Exibe o primeiro valor de dados encontrado no arquivo de dados.

Média Geométrica. A n-ésima raiz do produto dos valores de dados, em que n representa o número de casos.

Mediana Agrupada. Mediana que é calculada para dados que são codificados em grupos. Por exemplo, com dados de idade, se cada valor nos anos 30 for codificado como 35, cada valor nos anos 40 for codificado como 45, e assim por diante, a mediana agrupada será a média calculada a partir dos dados codificados.

Média Harmônica. Utilizada para estimar um tamanho médio do grupo quando os tamanhos da amostra nos grupos não são iguais. A média harmônica é o número total de amostras dividido pela soma dos recíprocos dos tamanhos da amostra.

Curtose. Uma medida da extensão até a qual as observações se agrupam em torno de um ponto central. Para a uma distribuição normal, o valor da estatística da curtose é zero. A curtose positiva indica que, com relação a uma distribuição normal, as observações são agrupadas mais ao centro da distribuição e têm rastros mais finos até os valores extremos da distribuição, no ponto em que os rastros da distribuição leptocúrtica são mais espessos com relação a uma distribuição normal. Já a curtose negativa indica que, com relação a uma distribuição normal, as observações são menos agrupadas e possuem rodapés mais espessos até os valores extremos da distribuição, no ponto em que os rastros da distribuição platykurtic são mais finos com relação a uma distribuição normal.

Last. Exibe o último valor de dados encontrado no arquivo de dados.

Máximo. O maior valor de uma variável numérica.

Média. Uma medida de tendência central. A média aritmética, a soma dividida pelo número de casos.

Median. O valor acima e abaixo do qual metade dos casos cai, o 50º percentil. Se houver um número par de casos, a mediana é a média dos dois casos intermediários quando estão classificados em ordem crescente ou decrescente. A mediana é uma medida da tendência central não sensível a valores excessivos (ao contrário da média, que pode ser afetada por alguns valores extremamente altos ou baixos).

Mínimo. O menor valor de uma variável numérica.

N. O número de casos (observações ou registros).

Percentual do N Total. Porcentagem do número total de casos em cada categoria.

Percentual da Soma Total. Porcentagem da soma total em cada categoria.

Intervalo. A diferença entre os valores maior e menor de uma variável numérica, o máximo menos o mínimo.

Assimetria. Uma medida da assimetria de uma distribuição. A distribuição normal é simétrica e tem um valor de assimetria de 0. Uma distribuição com assimetria positiva significativa tem um longo rodapé direito. Uma distribuição com assimetria negativa significativa tem um longo rodapé esquerdo. Como uma orientação, um valor de assimetria mais de duas vezes seu erro padrão é obtido para indicar uma partida de simetria.

Desvio padrão. Uma medida de dispersão em torno da média. Em uma distribuição normal, 68% dos casos estão dentro de um desvio padrão da média e 95% dos casos estão dentro de dois desvios padrão. Por exemplo, se a idade média for 45, com um desvio padrão de 10, 95% dos casos estariam entre 25 e 65 em uma distribuição normal.

Erro Padrão de Curtose. A razão da curtose com seu erro padrão pode ser utilizada como um teste de normalidade (ou seja, é possível rejeitar a normalidade se a razão for inferior a -2 ou superior a +2). Um valor positivo grande para curtose indica que os rodapés da distribuição são maiores do que aqueles de uma distribuição normal, e um valor negativo para curtose indica rodapés mais curtos (como aqueles de uma distribuição uniforme em forma de caixa).

Erro padrão de média. Uma medida do quanto o valor da média pode variar de amostra para amostra obtida da mesma distribuição. Ela pode ser utilizada para comparar aproximadamente a média observada com um valor hipotético (ou seja, é possível concluir que os dois valores serão diferentes se a razão da diferença com o erro padrão for inferior a -2 ou superior a +2).

Erro Padrão de Assimetria. A razão de assimetria com seu erro padrão pode ser utilizada como um teste de normalidade (ou seja, é possível rejeitar a normalidade se a razão for inferior a -2 ou superior a +2). Um valor positivo grande para assimetria indica um rodapé direito longo, e um valor negativo extremo indica um rodapé esquerdo longo.

Sum. A soma ou o total dos valores, em todos os casos com valores não omissos.

Variância. Uma medida de dispersão ao redor da média, igual à soma dos desvios quadrados da média dividido por um menor que o número de casos. A variância é medida em unidades que são o quadrado daquelas da própria variável.

Capítulo 7. Médias

O procedimento Médias calcula médias de subgrupo e estatísticas univariadas relacionadas para variáveis dependentes dentro das categorias de uma ou mais variáveis independentes. Opcionalmente, é possível obter uma análise de variância para um fator, eta e testes para linearidade.

Exemplo. Avalie a quantidade média de gordura absorvida por três tipos diferentes de óleo de cozinha, e execute uma análise de variância para um fator para ver se as médias diferem.

Estatísticas. Soma, número de casos, média, mediana, mediana agrupada, erro padrão da média, mínimo, máximo, intervalo, valor da variável da primeira categoria da variável de agrupamento, valor da variável da última categoria da variável de agrupamento, desvio padrão, variância, curtose, erro padrão da curtose, assimetria, erro padrão da assimetria, porcentagem da soma total, porcentagem do total de *N*, porcentagem da soma em, porcentagem de *N* em, média geométrica e média harmônica. As opções incluem análise de variância, eta, eta quadrado, e testes para linearidade *R* e R^2.

Considerações de Dados de Médias

Dados. As variáveis dependentes são quantitativas, e as variáveis independentes são categóricas. Os valores de variáveis categóricas podem ser numéricos ou sequência de caracteres.

Suposições. Algumas das estatísticas de subgrupo opcionais, como a média e o desvio padrão, baseiam-se em teoria normal e são apropriadas para variáveis quantitativas com distribuições simétricas. Estatísticas robustas, como a mediana, são apropriadas para variáveis quantitativas que podem ou não satisfazer a suposição de normalidade. A análise de variância é robusta para partidas da normalidade, mas os dados em cada célula devam ser simétricos. A análise de variância também supõe que os grupos vêm de populações com variâncias iguais. Para testar essa suposição, use o teste de homogeneidade de variância de Levene, disponível no procedimento One-Way ANOVA.

Para Obter Médias de Subgrupo

1. Nos menus, escolha:

   Analisar > **Comparar Médias** > **Médias...**
2. Selecione uma ou mais variáveis dependentes.
3. Use um dos seguintes métodos para selecionar variáveis independentes categóricas:
   - Selecione uma ou mais variáveis independentes. Resultados separados são exibidos para cada variável independente.
   - Selecione uma ou mais estratos para variáveis independentes. Cada estrato subdivide ainda mais a amostra. Se você tiver uma variável independente na Camada 1 e uma variável independente na Camada 2, os resultados serão exibidos em uma tabela cruzada, ao invés de tabelas separadas para cada variável independente.
4. Opcionalmente, clique em **Opções** para estatísticas opcionais, uma análise de tabela de variância, eta, eta quadrado, *R* e R^2.

Opções de Médias

É possível escolher uma ou mais das seguintes estatísticas de subgrupo para as variáveis dentro de cada categoria de cada variável de agrupamento: soma, número de casos, média, mediana, mediana agrupada, erro padrão da média, mínimo, máximo, intervalo, valor da variável da primeira categoria da variável de agrupamento, valor da variável da última categoria da variável de agrupamento, desvio padrão, variância, curtose, erro padrão da curtose, assimetria, erro padrão da assimetria, porcentagem da soma total, porcentagem do total de *N*, porcentagem da soma em, porcentagem de *N* em, média geométrica e

média harmônica. É possível alterar a ordem na qual as estatísticas de subgrupo aparecem. A ordem na qual as estatísticas aparecem na lista de Estatísticas de Célula é a ordem na qual elas são exibidas na saída. As estatísticas de sumarização também são exibidas para cada variável em todas as categorias.

First. Exibe o primeiro valor de dados encontrado no arquivo de dados.

Média Geométrica. A n-ésima raiz do produto dos valores de dados, em que n representa o número de casos.

Mediana Agrupada. Mediana que é calculada para dados que são codificados em grupos. Por exemplo, com dados de idade, se cada valor nos anos 30 for codificado como 35, cada valor nos anos 40 for codificado como 45, e assim por diante, a mediana agrupada será a média calculada a partir dos dados codificados.

Média Harmônica. Utilizada para estimar um tamanho médio do grupo quando os tamanhos da amostra nos grupos não são iguais. A média harmônica é o número total de amostras dividido pela soma dos recíprocos dos tamanhos da amostra.

Curtose. Uma medida da extensão até a qual as observações se agrupam em torno de um ponto central. Para a uma distribuição normal, o valor da estatística da curtose é zero. A curtose positiva indica que, com relação a uma distribuição normal, as observações são agrupadas mais ao centro da distribuição e têm rastros mais finos até os valores extremos da distribuição, no ponto em que os rastros da distribuição leptocúrtica são mais espessos com relação a uma distribuição normal. Já a curtose negativa indica que, com relação a uma distribuição normal, as observações são menos agrupadas e possuem rodapés mais espessos até os valores extremos da distribuição, no ponto em que os rastros da distribuição platykurtic são mais finos com relação a uma distribuição normal.

Last. Exibe o último valor de dados encontrado no arquivo de dados.

Máximo. O maior valor de uma variável numérica.

Média. Uma medida de tendência central. A média aritmética, a soma dividida pelo número de casos.

Median. O valor acima e abaixo do qual metade dos casos cai, o 50º percentil. Se houver um número par de casos, a mediana é a média dos dois casos intermediários quando estão classificados em ordem crescente ou decrescente. A mediana é uma medida da tendência central não sensível a valores excessivos (ao contrário da média, que pode ser afetada por alguns valores extremamente altos ou baixos).

Mínimo. O menor valor de uma variável numérica.

N. O número de casos (observações ou registros).

Percentual de N total. Porcentagem do número total de casos em cada categoria.

Percentual da soma total. Porcentagem da soma total em cada categoria.

Intervalo. A diferença entre os valores maior e menor de uma variável numérica, o máximo menos o mínimo.

Assimetria. Uma medida da assimetria de uma distribuição. A distribuição normal é simétrica e tem um valor de assimetria de 0. Uma distribuição com assimetria positiva significativa tem um longo rodapé direito. Uma distribuição com assimetria negativa significativa tem um longo rodapé esquerdo. Como uma orientação, um valor de assimetria mais de duas vezes seu erro padrão é obtido para indicar uma partida de simetria.

Desvio padrão. Uma medida de dispersão em torno da média. Em uma distribuição normal, 68% dos casos estão dentro de um desvio padrão da média e 95% dos casos estão dentro de dois desvios padrão. Por exemplo, se a idade média for 45, com um desvio padrão de 10, 95% dos casos estariam entre 25 e 65 em uma distribuição normal.

Erro Padrão de Curtose. A razão da curtose com seu erro padrão pode ser utilizada como um teste de normalidade (ou seja, é possível rejeitar a normalidade se a razão for inferior a -2 ou superior a +2). Um valor positivo grande para curtose indica que os rodapés da distribuição são maiores do que aqueles de uma distribuição normal, e um valor negativo para curtose indica rodapés mais curtos (como aqueles de uma distribuição uniforme em forma de caixa).

Erro padrão de média. Uma medida do quanto o valor da média pode variar de amostra para amostra obtida da mesma distribuição. Ela pode ser utilizada para comparar aproximadamente a média observada com um valor hipotético (ou seja, é possível concluir que os dois valores serão diferentes se a razão da diferença com o erro padrão for inferior a -2 ou superior a +2).

Erro Padrão de Assimetria. A razão de assimetria com seu erro padrão pode ser utilizada como um teste de normalidade (ou seja, é possível rejeitar a normalidade se a razão for inferior a -2 ou superior a +2). Um valor positivo grande para assimetria indica um rodapé direito longo, e um valor negativo extremo indica um rodapé esquerdo longo.

Sum. A soma ou o total dos valores, em todos os casos com valores não omissos.

Variância. Uma medida de dispersão ao redor da média, igual à soma dos desvios quadrados da média dividido por um menor que o número de casos. A variância é medida em unidades que são o quadrado daquelas da própria variável.

Estatísticas para a primeira estrato

Tabela anova e o eta. Exibe uma tabela de análise de variância unidirecional e calcula o eta e o eta quadrado (medidas de associação) para cada variável independente na primeira camada.

Teste para linearidade. Calcula a soma dos quadrados, os graus de liberdade e o quadrado médio associados aos componentes lineares e não lineares, bem como a razão de F, R e R-quadrado. A linearidade não será calculada se a variável independente for uma sequência de caracteres curta.

Capítulo 8. Cubos OLAP

O procedimento Cubos do OLAP (Online Analytical Processing) calcula totais, médias e outras estatísticas univariadas para variáveis de sumarização contínuas dentro de categorias de uma ou mais variáveis de agrupamento categóricas. Uma estrato separada na tabela é criada para cada categoria de cada variável de agrupamento.

Exemplo. Vendas total e média para diferentes regiões e linhas de produtos dentro de regiões.

Estatísticas. Soma, número de casos, média, mediana, mediana agrupada, erro padrão da média, mínimo, máximo, intervalo, valor da variável da primeira categoria da variável de agrupamento, valor da variável da última categoria da variável de agrupamento, desvio padrão, variância, curtose, erro padrão da curtose, assimetria, erro padrão da assimetria, porcentagem do total de casos, porcentagem da soma total, porcentagem do total de casos em variáveis de agrupamento, porcentagem da soma total em variáveis de agrupamento, média geométrica e a média harmônica.

Considerações de Dados de Cubos do OLAP

Dados. As variáveis de sumarização são quantitativas (variáveis contínuas medidas em uma escala de intervalo ou de razão) e as variáveis de agrupamento são categóricas. Os valores de variáveis categóricas podem ser numéricos ou sequência de caracteres.

Suposições. Algumas das estatísticas de subgrupo opcionais, como a média e o desvio padrão, baseiam-se em teoria normal e são apropriadas para variáveis quantitativas com distribuições simétricas. Estatísticas robustas, como a mediana e o intervalo, são apropriadas para variáveis quantitativas que podem ou não satisfazer a suposição de normalidade.

Para Obter Cubos OLAP

1. Nos menus, escolha:

   Analisar > **Relatórios** > **Cubos OLAP...**
2. Selecione uma ou mais variáveis de sumarização contínuas.
3. Selecione uma ou mais variáveis de agrupamento categóricas.

Opcionalmente:

- Selecione estatísticas básicas diferentes (clique em **Estatísticas**). Deve-se selecionar uma ou mais variáveis de agrupamento antes de poder selecionar as estatísticas básicas.
- Calcule diferenças entre os pares de variáveis e pares de grupos que estiverem definidos por uma variável de agrupamento (clique em **Diferenças**).
- Crie títulos de tabela customizados (clique em **Título**).
- Oculte contagens que forem menores que um número inteiro especificado. Valores ocultos serão exibidos como **<N**, em que **N** é o número inteiro especificado. O número inteiro especificado deve ser maior ou igual a 2.

Estatísticas de Cubos OLAP

É possível escolher uma ou mais das seguintes estatísticas de subgrupo para as variáveis de sumarização dentro de cada categoria de cada variável de agrupamento: soma, número de casos, média, mediana, mediana agrupada, erro padrão da média, mínimo, máximo, intervalo, valor da variável da primeira categoria da variável de agrupamento, valor da variável da última categoria da variável de agrupamento, desvio padrão, variância, curtose, erro padrão da curtose, assimetria, erro padrão da assimetria,

porcentagem do total de casos, porcentagem da soma total, porcentagem do total de casos em variáveis de agrupamento, porcentagem da soma total em variáveis de agrupamento, média geométrica e a média harmônica.

É possível alterar a ordem na qual as estatísticas de subgrupo aparecem. A ordem na qual as estatísticas aparecem na lista de Estatísticas de Célula é a ordem na qual elas são exibidas na saída. As estatísticas de sumarização também são exibidas para cada variável em todas as categorias.

First. Exibe o primeiro valor de dados encontrado no arquivo de dados.

Média Geométrica. A n-ésima raiz do produto dos valores de dados, em que n representa o número de casos.

Mediana Agrupada. Mediana que é calculada para dados que são codificados em grupos. Por exemplo, com dados de idade, se cada valor nos anos 30 for codificado como 35, cada valor nos anos 40 for codificado como 45, e assim por diante, a mediana agrupada será a média calculada a partir dos dados codificados.

Média Harmônica. Utilizada para estimar um tamanho médio do grupo quando os tamanhos da amostra nos grupos não são iguais. A média harmônica é o número total de amostras dividido pela soma dos recíprocos dos tamanhos da amostra.

Curtose. Uma medida da extensão até a qual as observações se agrupam em torno de um ponto central. Para a uma distribuição normal, o valor da estatística da curtose é zero. A curtose positiva indica que, com relação a uma distribuição normal, as observações são agrupadas mais ao centro da distribuição e têm rastros mais finos até os valores extremos da distribuição, no ponto em que os rastros da distribuição leptocúrtica são mais espessos com relação a uma distribuição normal. Já a curtose negativa indica que, com relação a uma distribuição normal, as observações são menos agrupadas e possuem rodapés mais espessos até os valores extremos da distribuição, no ponto em que os rastros da distribuição platykurtic são mais finos com relação a uma distribuição normal.

Last. Exibe o último valor de dados encontrado no arquivo de dados.

Máximo. O maior valor de uma variável numérica.

Média. Uma medida de tendência central. A média aritmética, a soma dividida pelo número de casos.

Median. O valor acima e abaixo do qual metade dos casos cai, o 50º percentil. Se houver um número par de casos, a mediana é a média dos dois casos intermediários quando estão classificados em ordem crescente ou decrescente. A mediana é uma medida da tendência central não sensível a valores excessivos (ao contrário da média, que pode ser afetada por alguns valores extremamente altos ou baixos).

Mínimo. O menor valor de uma variável numérica.

N. O número de casos (observações ou registros).

Percentual de N em. Porcentagem do número de casos para a variável de agrupamento especificada dentro das categorias de outras variáveis de agrupamento. Se você tiver apenas uma variável de agrupamento, esse valor será idêntico à porcentagem do número total de casos.

Percentual da Soma em. Porcentagem da soma para a variável de agrupamento especificada dentro das categorias de outras variáveis de agrupamento. Se você tiver apenas uma variável de agrupamento, esse valor será idêntico à porcentagem da soma total.

Percentual do N Total. Porcentagem do número total de casos em cada categoria.

Percentual da Soma Total. Porcentagem da soma total em cada categoria.

Intervalo. A diferença entre os valores maior e menor de uma variável numérica, o máximo menos o mínimo.

Assimetria. Uma medida da assimetria de uma distribuição. A distribuição normal é simétrica e tem um valor de assimetria de 0. Uma distribuição com assimetria positiva significativa tem um longo rodapé direito. Uma distribuição com assimetria negativa significativa tem um longo rodapé esquerdo. Como uma orientação, um valor de assimetria mais de duas vezes seu erro padrão é obtido para indicar uma partida de simetria.

Desvio padrão. Uma medida de dispersão em torno da média. Em uma distribuição normal, 68% dos casos estão dentro de um desvio padrão da média e 95% dos casos estão dentro de dois desvios padrão. Por exemplo, se a idade média for 45, com um desvio padrão de 10, 95% dos casos estariam entre 25 e 65 em uma distribuição normal.

Erro Padrão de Curtose. A razão da curtose com seu erro padrão pode ser utilizada como um teste de normalidade (ou seja, é possível rejeitar a normalidade se a razão for inferior a -2 ou superior a +2). Um valor positivo grande para curtose indica que os rodapés da distribuição são maiores do que aqueles de uma distribuição normal, e um valor negativo para curtose indica rodapés mais curtos (como aqueles de uma distribuição uniforme em forma de caixa).

Erro padrão de média. Uma medida do quanto o valor da média pode variar de amostra para amostra obtida da mesma distribuição. Ela pode ser utilizada para comparar aproximadamente a média observada com um valor hipotético (ou seja, é possível concluir que os dois valores serão diferentes se a razão da diferença com o erro padrão for inferior a -2 ou superior a +2).

Erro Padrão de Assimetria. A razão de assimetria com seu erro padrão pode ser utilizada como um teste de normalidade (ou seja, é possível rejeitar a normalidade se a razão for inferior a -2 ou superior a +2). Um valor positivo grande para assimetria indica um rodapé direito longo, e um valor negativo extremo indica um rodapé esquerdo longo.

Sum. A soma ou o total dos valores, em todos os casos com valores não omissos.

Variância. Uma medida de dispersão ao redor da média, igual à soma dos desvios quadrados da média dividido por um menor que o número de casos. A variância é medida em unidades que são o quadrado daquelas da própria variável.

Diferenças de Cubos OLAP

Esta caixa de diálogo permite calcular as diferenças de porcentagem e aritméticas entre variáveis de sumarização ou entre grupos que são definidos por uma variável de agrupamento. As diferenças são calculadas para todas as medidas que estiverem selecionadas na caixa de diálogo Estatísticas de Cubos OLAP.

Diferenças entre Variáveis. Calcula diferenças entre os pares de variáveis. Os valores de estatísticas básicas para a segunda variável (a variável Minus) em cada par são subtraídos dos valores de estatísticas básicas para a primeira variável no par. Para obter as diferenças de porcentagem, o valor da variável de sumarização para a variável Minus é usado como o denominador. Deve-se selecionar pelo menos duas variáveis de sumarização na caixa de diálogo principal antes de poder especificar as diferenças entre as variáveis.

Diferenças entre Grupos de Casos. Calcula diferenças entre pares de grupos definidos por uma variável de agrupamento. Os valores de estatísticas básicas para a segunda categoria (a categoria Minus) são subtraídos dos valores de estatísticas básicas para a primeira categoria no par. As diferenças de

porcentagem utilizam o valor da estatística de sumarização para a categoria Minus como o denominador. Deve-se selecionar uma ou mais variáveis de agrupamento na caixa de diálogo principal antes de poder especificar diferenças entre os grupos.

Título de Cubos OLAP

É possível alterar o título de sua saída ou incluir uma legenda que aparecerá abaixo da tabela de saída. Também é possível controlar a quebra de linha em títulos e legendas ao digitar `\n` onde desejar inserir uma quebra de linha no texto.

Capítulo 9. Teste t

Testes t

Três tipos de testes *t* estão disponíveis:

Teste t de amostras independentes (teste T de duas amostras). Compara o meio de uma variável para dois grupos de casos. Estatísticas descritivas para cada grupo e o teste de Levene para igualdade de variações são fornecidos, bem como os valores de *t* de variação igual e desigual e um intervalo de confiança de 95% para a diferença em média.

Teste t de amostras pairwise teste t dependente). Compara a média de duas variáveis para um único grupo. Este teste também é destinado a pares correspondidos ou designs de estudo de caso-controle. A saída inclui estatísticas descritivas para as variáveis de teste, a correlação entre as variáveis, as estatísticas descritivas para diferenças pairwise, o teste *t* e um intervalo de confiança de 95%.

Teste t de uma amostra. Compara a média de uma variável com um valor conhecido ou hipotético. Estatísticas descritivas para as variáveis de teste são exibidas junto com o teste *t*. Um intervalo de confiança de 95% para a diferença entre a média da variável do teste e o valor de teste hipotético é parte da saída padrão.

Teste-T de amostras independentes

O procedimento Teste T de Amostras Independentes compara médias de dois grupos de casos. Idealmente, para este teste, os sujeitos devem ser designados aleatoriamente para dois grupos, de forma que qualquer diferença na resposta será devido ao tratamento (ou à falta de tratamento) e não a outros fatores. Esse não será o caso se você comparar uma renda média para homens e mulheres. Uma pessoa não é designada aleatoriamente para ser homem ou mulher. Nessas situações, é necessário assegurar-se de que as diferenças nos outros fatores não venham mascarar ou aprimorar uma diferença significativa nas médias. As diferenças na renda média podem ser influenciadas por fatores como educação (e não somente pelo sexo).

Exemplo. Pacientes com pressão alta são designados aleatoriamente para um grupo placebo e para um grupo de tratamento. Os sujeitos no grupo placebo recebem um comprimido inativo, e os sujeitos no grupo de tratamento recebem uma nova droga que se espera que diminua a pressão arterial. Após dois meses de tratamento, o teste *t* de duas amostras é utilizado para comparar a pressão arterial média do grupo placebo e do grupo de tratamento. Cada paciente é avaliado uma vez e pertence a um grupo.

Estatísticas. Para cada variável: tamanho da amostra, média, desvio padrão e erro padrão da média. Para a diferença nas médias: média, erro padrão e intervalo de confiança (é possível especificar o nível de confiança). Testes: teste de Levene de igualdade de variâncias e testes *t* de variâncias agrupadas e variâncias separadas para igualdade de médias.

Considerações de Dados do Teste T de Amostras Independentes

Dados. Os valores da variável quantitativa de interesse estão em uma única coluna no arquivo de dados. O procedimento usa uma variável de agrupamento com dois valores para separar os casos em dois grupos. A variável de agrupamento pode ser numérica (valores como 1 e 2 ou 6,25 e 12,5) ou uma sequência curta (como *sim* e *não*). Como alternativa, é possível utilizar uma variável quantitativa, como *idade*, para dividir os casos em dois grupos especificando um ponto de corte (o ponto de corte 21 divide *idade* em um grupo sub-21 e um grupo 21 e acima).

Suposições. Para o teste *t* de variância igual, as observações devem ser amostras aleatórias independentes das distribuições normais com a mesma variância da população. Para o teste *t* de variância desigual, as observações devem ser amostras aleatórias independentes das distribuições normais. O teste *t* de duas amostras é bastante robusto para partidas da normalidade. Ao verificar distribuições graficamente, veja que elas são simétricas e não possuem nenhum valor discrepante.

Para Obter um Teste T de Amostras Independentes

1. Nos menus, escolha:

   Analisar > **Comparar Médias** > **Teste T de Amostras Independentes ...**
2. Selecione uma ou mais variáveis de teste quantitativas. Um teste *t* separado é calculado para cada variável.
3. Selecione uma variável de agrupamento única e, em seguida, clique em **Definir Grupos** para especificar dois códigos para os grupos que deseja comparar.
4. Opcionalmente, clique em **Opções** para controlar o tratamento de dados omissos e o nível do intervalo de confiança.

Definir Grupos para Teste-T de Amostras Independentes

Para variáveis de agrupamento numéricas, defina os dois grupos para o teste *t* ao especificar dois valores ou um ponto de corte:

- **Utilizar valores especificados.** Insira um valor para o Grupo 1 e outro valor para o Grupo 2. Casos com quaisquer outros valores serão excluídos da análise. Os números não precisam ser inteiros (por exemplo, 6,25 e 12,5 são válidos).
- **Ponto de Corte.** Insira um número que divide os valores da variável de agrupamento em dois conjuntos. Todos os casos com valores que forem menores que o ponto de corte formam um grupo, e os casos com valores que forem maiores ou iguais ao ponto de corte formam o outro grupo.

Para variáveis de agrupamento de sequência de caracteres, insira uma sequência de caracteres para o Grupo 1 e outro valor para o Grupo 2, como *sim* e *não*. Casos com outras sequências de caracteres são excluídos da análise.

Opções de Teste T de Amostras Independentes

Intervalo de Confiança. Por padrão, um intervalo de confiança de 95% para a diferença nas médias é exibido. Insira um valor entre 1 e 99 para solicitar um nível de confiança diferente.

Valores omissos. Quando testar várias variáveis, e os dados estiverem omissos para uma ou mais variáveis, é possível informar ao procedimento quais casos devem ser incluídos (ou excluídos).

- **Excluir análise de casos por análise.** Cada teste *t* utiliza todos os casos que possuírem dados válidos para as variáveis testadas. Os tamanhos de amostra podem variar de teste para teste.
- **Excluir listwise dos casos.** Cada teste *t* utiliza apenas os casos que possuírem dados válidos para todas as variáveis que são utilizadas nos testes *t* solicitados. O tamanho da amostra é constante nos testes.

Teste-T de amostras em pares

O procedimento de Teste T de Amostras Emparelhadas compara as médias de duas variáveis para um único grupo. O procedimento calcula as diferenças entre os valores das duas variáveis para cada caso e testa se a média difere de 0.

Exemplo. Em um estudo sobre a pressão arterial alta, todos os pacientes são medidos no início do estudo, dado um tratamento, e medidos novamente. Assim, cada sujeito possui duas medidas, geralmente chamadas de medidas *antes* e *após*. Um design alternativo para o qual este teste é utilizado é um estudo de correspondência de pares e de caso de controle, em que cada registro no arquivo de dados contém a

resposta para o paciente e também para seu sujeito de controle correspondente. Em um estudo de pressão arterial, os pacientes e os controles podem ser correspondidos por idade (um paciente de 75 anos de idade com um membro do grupo de controle também de 75 anos).

Estatísticas. Para cada variável: média, tamanho da amostra, desvio padrão e erro padrão da média. Para cada par de variáveis: correlação, diferença média nas médias, teste *t* e intervalo de confiança para a diferença média (é possível especificar o nível de confiança). Desvio padrão e o erro padrão da diferença média.

Considerações de Dados do Teste T de Amostras Emparelhadas

Dados. Para cada teste emparelhado, especifique duas variáveis quantitativas (nível de medição de intervalo ou nível de medição de razão). Para um estudo de correspondência de pares e de caso de controle, a resposta para cada sujeito de teste e seu sujeito de controle correspondente deve estar no mesmo caso no arquivo de dados.

Suposições. Observações para cada par devem ser feitas nas mesmas condições. A diferença média deve ser normalmente distribuída. Variâncias de cada variável podem ser iguais ou desiguais.

Para Obter um Teste T de Amostras Emparelhadas

1. Nos menus, escolha:

   Analisar > **Comparar Médias** > **Teste T de Amostras Emparelhadas ...**
2. Selecione um ou mais pares de variáveis
3. Opcionalmente, clique em **Opções** para controlar o tratamento de dados omissos e o nível do intervalo de confiança.

Opções de Teste T de Amostras Emparelhadas

Intervalo de Confiança. Por padrão, um intervalo de confiança de 95% para a diferença nas médias é exibido. Insira um valor entre 1 e 99 para solicitar um nível de confiança diferente.

Valores omissos. Quando testar várias variáveis, e os dados estiverem omissos para uma ou mais variáveis, é possível informar ao procedimento quais casos devem ser incluídos (ou excluídos):

- **Excluir análise de casos por análise.** Cada teste *t* utiliza todos os casos que possuírem dados válidos para o par de variáveis testado. Os tamanhos de amostra podem variar de teste para teste.
- **Excluir listwise dos casos.** Cada teste *t* utiliza apenas os casos que possuírem dados válidos para todos os pares de variáveis testados. O tamanho da amostra é constante nos testes.

Recursos adicionais do comando T-TEST

O idioma da sintaxe de comando também permite:

- Produzir ambos os teste t - de uma amostra e de amostras independentes - executando um único comando.
- Testar uma variável com relação a cada variável em uma lista em um teste t emparelhado (com o subcomando `PAIRS`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Teste-T de uma amostra

O procedimento Teste T de Uma Amostra testa se a média de uma variável única difere de uma constante especificada.

Exemplos. Um pesquisador pode querer testar se o escore médio de QI para um grupo de estudantes difere de 100. Ou um fabricante de cereais pode tirar uma amostra das caixas da linha de produção e verificar se o peso médio das amostras tem uma diferença de 1,3 libras no nível de confiança de 95%.

Estatísticas. Para cada variável de teste: média, desvio padrão e erro padrão da média. A diferença média entre cada valor de dados e o valor de teste hipotético, um teste *t* que confirma que essa diferença é 0, e um intervalo de confiança para esta diferença (é possível especificar o nível de confiança).

Considerações de Dados do Teste T de Uma Amostra

Dados. Para testar os valores de uma variável quantitativa com relação a um valor de teste hipotético, escolha uma variável quantitativa e insira um valor de teste hipotético.

Suposições. Esse teste supõe que os dados são normalmente distribuídos; entretanto, esse teste é bastante robusto para partidas da normalidade.

Para Obter um Teste T de Uma Amostra

1. Nos menus, escolha:

   Analisar > **Comparar Médias** > **Teste T de Uma Amostra ...**
2. Selecione uma ou mais variáveis a serem testadas com relação ao mesmo valor hipotético.
3. Insira um valor de teste numérico com relação ao qual cada média da amostra é comparada.
4. Opcionalmente, clique em **Opções** para controlar o tratamento de dados omissos e o nível do intervalo de confiança.

Opções de Teste T de Uma Amostra

Intervalo de Confiança. Por padrão, um intervalo de confiança de 95% para a diferença entre a média e o valor de teste hipotético é exibido. Insira um valor entre 1 e 99 para solicitar um nível de confiança diferente.

Valores omissos. Quando testar várias variáveis, e os dados estiverem omissos para uma ou mais variáveis, é possível informar ao procedimento quais casos devem ser incluídos (ou excluídos).

- **Excluir análise de casos por análise.** Cada teste *t* utiliza todos os casos que possuírem dados válidos para a variável testada. Os tamanhos de amostra podem variar de teste para teste.
- **Excluir listwise dos casos.** Cada teste *t* utiliza apenas os casos que possuírem dados válidos para todas as variáveis que são utilizadas em qualquer um dos testes *t* solicitados. O tamanho da amostra é constante nos testes.

Recursos adicionais do comando T-TEST

O idioma da sintaxe de comando também permite:

- Produzir ambos os teste t - de uma amostra e de amostras independentes - executando um único comando.
- Testar uma variável com relação a cada variável em uma lista em um teste t emparelhado (com o subcomando PAIRS).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Recursos adicionais do comando T-TEST

O idioma da sintaxe de comando também permite:

- Produzir ambos os teste t - de uma amostra e de amostras independentes - executando um único comando.

- Testar uma variável com relação a cada variável em uma lista em um teste t emparelhado (com o subcomando PAIRS).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 10. One-Way ANOVA

O procedimento One-Way ANOVA produz uma análise de variância por um fator para uma variável dependente quantitativa por uma única variável (independente) de fator. A análise de variância é utilizada para testar a hipótese de que várias médias são iguais. Esta técnica é uma extensão do teste *t* de duas amostras.

Além de determinar que existem diferenças entre as médias, talvez você queira saber quais médias diferem. Há dois tipos de testes para comparar médias: contrastes a priori e testes post hoc. Contrastes são testes configurados *antes de* executar o experimento, e testes post hoc são executados *após* o experimento ter sido conduzido. Também é possível testar tendências entre as categorias.

Exemplo. Doces como sonhos de padarias absorvem várias quantias de gordura quando são preparados. Um experimento é configurado envolvendo três tipos de gordura: óleo de amendoim, óleo de milho e banha de porco. O óleo de amendoim e o óleo de milho são gorduras insaturadas, e a banha de porco é uma gordura saturada. Além de determinar se a quantia de gordura absorvida depende do tipo de gordura utilizada, é possível configurar um contraste a priori para determinar se a quantia de absorção de gordura difere para gorduras saturadas e insaturadas.

Estatísticas. Para cada grupo: número de casos, média, desvio padrão, erro padrão da média, mínimo, máximo e o intervalo de confiança de 95% para a média. Teste de Levene para a homogeneidade de variância, tabela de análise de variância e testes robusto da igualdade das médias para cada variável dependente, contrastes a priori especificados pelo usuário e testes de intervalo post hoc e diversas comparações: Bonferroni, Sidak, diferença significativa honesta de Tukey, GT2 de Hochberg, Gabriel, Dunnett, teste *F* de Ryan-Einot-Gabriel-Welsch (*F* de R-E-G-W), teste de intervalo de Ryan-Einot-Gabriel-Welsch (*Q* de R-E-G-W), T2 de Tamhane, T3 de Dunnett, Games-Howell, *C* de Dunnett, teste de amplitude múltipla de Duncan, Student-Newman-Keuls (S-N-K), *b* de Tukey, Waller-Duncan, Scheffé e diferença menos significativa.

Considerações de Dados de One-Way ANOVA

Dados. Os valores da variável de fator devem ser números inteiros, e a variável dependente deverá ser quantitativa (nível de medição de intervalo).

Suposições. Cada grupo é uma amostra aleatória independente de uma população normal. A análise de variância é robusta para partidas da normalidade, embora os dados devam ser simétricos. Os grupos devem vir de populações com variâncias iguais. Para testar essa suposição, use o teste de homogeneidade de variância de Levene.

Para Obter uma Análise de Variância para Um Fator

1. Nos menus, escolha:
   Analisar > **Comparar Médias** > **One-Way ANOVA...**
2. Selecione uma ou mais variáveis dependentes.
3. Selecione uma variável de fator independente única.

Contrastes One-Way ANOVA

É possível particionar as somas dos quadrados entre grupos em componentes de tendência ou especificar contrastes a priori.

Polinomial. Particiona as somas dos quadrados entre grupos em componentes de tendência. É possível testar para uma tendência da variável dependente entre os níveis ordenados da variável de fator. Por exemplo, é possível testar para uma tendência linear (aumento ou diminuição) de salário entre os níveis ordenados do mais alto grau obtido.

- **Grau.** É possível escolher um polinomial de 1º, 2º, 3º, 4º ou 5º grau.

Coeficientes. Contrastes a priori especificados pelo usuário a serem testados pela estatística *t*. Insira um coeficiente para cada grupo (categoria) da variável de fator e clique em **Incluir** após cada entrada. Cada novo valor é incluído na parte inferior da lista de coeficientes. Para especificar conjuntos de contrastes adicionais, clique em **Avançar**. Utilize **Avançar** e **Anterior** para mover entre os conjuntos de contrastes.

A ordem dos coeficientes é importante porque ela corresponde à ordem crescente dos valores da categoria da variável de fator. O primeiro coeficiente na lista corresponde ao menor valor do grupo da variável de fator, e o último coeficiente corresponde ao valor mais alto. Por exemplo, se houver seis categorias da variável de fator, os coeficientes -1, 0, 0, 0, 0,5, e 0,5 contrastam o primeiro grupo com os quinto e sexto grupos. Para a maioria das aplicações, os coeficientes devem somar 0. Conjuntos que não somam 0 também podem ser utilizados, mas uma mensagem de aviso é exibida.

Testes Post Hoc do One-Way ANOVA

Após determinar que diferenças existem entre as médias, testes de intervalo post hoc e comparações múltiplas entre pares podem determinar quais médias diferem. Os testes de intervalo identificam subconjuntos homogêneos de médias que não são diferentes uns dos outros. Comparações múltiplas entre pares testam a diferença entre cada par de médias e geram uma matriz em que os asteriscos indicam médias de grupo significativamente diferentes em um nível alfa de 0,05.

Variâncias iguais presumidas

O teste da diferença significativa honesta de Tukey, GT2 de Hochberg, Gabriel e Scheffé são testes de comparação múltipla e também testes de intervalo. Outros testes de intervalo disponíveis são *b* de Tukey, S-N-K (Student-Newman-Keuls), Duncan, *F* de R-E-G-W (teste *F* de Ryan-Einot-Gabriel-Welsch), *Q* de R-E-G-W (teste de intervalo de Ryan-Einot-Gabriel-Welsch) e Waller-Duncan. Testes de comparação múltipla disponíveis são Bonferroni, teste da diferença significativa honesta de Tukey, Sidak, Gabriel, Hochberg, Dunnett, Scheffé, e LSD (diferença menos significativa).

- *LSD*. Utiliza testes t para executar todas as comparações pairwise entre as médias de grupo. Nenhum ajustamento é feito na taxa de erros para comparações múltiplas.
- *Bonferroni*. Utiliza testes t para executar comparações pairwise entre as médias de grupo, e também controla a taxa de erro geral ao configurar a taxa de erro de cada teste com a taxa de erro entre experimentos dividido pelo número total de testes. Portanto, o nível de significância observado é ajustado para o fato de que diversas comparações estão sendo feitas.
- *Sidak*. Teste de comparação múltipla entre pares com base em uma estatística t. O Sidak ajusta o nível de significância para comparações múltiplas e fornece limites mais apertados que Bonferroni.
- *Scheffe*. Executa comparações pairwise conjuntas simultâneas para todas as combinações entre pares possíveis de médias. Utiliza a distribuição F de amostragem. Pode ser utilizado para examinar todas as combinações lineares possíveis das médias de grupo, não apenas comparações pairwise.
- *R-E-G-W F*. Procedimento stepdown múltiplo de Ryan-Einot-Gabriel-Welsch com base em um teste F.
- *R-E-G-W Q*. Procedimento stepdown múltiplo de Ryan-Einot-Gabriel-Welsch com base em um intervalo estudentizado.
- *S-N-K*. Faz todas as comparações pairwise entre as médias usando a distribuição de intervalo estudentizado. Com tamanhos de amostra iguais, ele também compara pares de médias em subconjuntos homogêneos, utilizando um procedimento stepwise. As médias são ordenadas da mais alta para a mais baixa, e as diferenças extremas são testadas em primeiro lugar.

- *Tukey*. Utiliza a estatística de intervalo Estudentizado para fazer todas as comparações pairwise entre os grupos. Configura a taxa de erros entre os experimentos na taxa de erros da coleção de todas as comparações pairwise.
- *b de Tukey*. Utiliza a distribuição de intervalo Estudentizado para fazer comparações pairwise entre os grupos. O valor crítico é a média do valor correspondente para o teste da diferença significativa honesta de Tukey e o Student-Newman-Keuls.
- *Duncan*. Faz comparações pairwise utilizando uma ordem de comparações stepwise idêntica à ordem utilizada pelo teste Student-Newman-Keuls, porém configura um nível de proteção para a taxa de erro da coleção de testes ao invés de configurar para uma taxa de erro de testes individuais. Utiliza a amplitude estatística estudentizada.
- *GT2 de Hochberg*. Teste múltiplo de comparação e intervalo que utiliza o módulo máximo estudentizado. Semelhante ao teste da diferença significativa honesta de Tukey.
- *Gabriel*. Teste de comparação pairwise que usa o módulo máximo estudentizado e geralmente é mais poderoso que o GT2 de Hochberg quando os tamanhos de células são desiguais. O teste de Gabriel pode ser liberal quando os tamanhos de células variam grandiosamente.
- *Waller-Duncan*. Teste de comparação múltipla com base em uma estatística t que utiliza abordagem Bayesiana.
- *Dunnett*. Teste t de comparação de diversos pares que compara um conjunto de tratamentos com relação a uma média de controle única. A última categoria é a categoria de controle padrão. Como alternativa, é possível escolher a primeira categoria. **Bilateral** testa se a média em qualquer nível (exceto a categoria de controle) do fator não é igual a da categoria de controle. **< Controle** testa se a média em qualquer nível do fator é menor do que a da categoria de controle. **> Controle** testa se a média em qualquer nível do fator é maior do que a da categoria de controle.

Variâncias iguais não presumidas

Testes de comparação múltipla não consideram que variâncias iguais sejam T2 de Tamhane, T3 de Dunnett, Games-Howell e *C* de Dunnett.

- *T2 de Tamhane*. Teste de comparação pairwise conservador com base em um teste t. Este teste é apropriado quando as variâncias são desiguais.
- *T3 de Dunnett*. Teste de comparação pairwise com base no Módulo máximo estudentizado. Este teste é apropriado quando as variâncias são desiguais.
- *Games-Howell*. Teste de comparação pairwise que às vezes é liberal. Este teste é apropriado quando as variâncias são desiguais.
- *C de Dunnett*. Teste de comparação pairwise com base no intervalo estudentizado. Este teste é apropriado quando as variâncias são desiguais.

Nota: Talvez você ache mais fácil interpretar a saída de testes post hoc se cancelar a seleção de **Ocultar linhas e colunas vazias** na caixa de diálogo Propriedades da Tabela (em uma tabela dinâmica ativada, escolha **Propriedades da Tabela** no menu Formatar).

Opções de One-Way ANOVA

Estatísticas. Escolha um ou mais dos seguintes:

- **Descritivo.** Calcula o número de casos, a média, o desvio padrão, o erro padrão da média, mínimo, máximo, e intervalos de confiança de 95% para cada variável dependente de cada grupo.
- **Efeitos fixos e aleatórios.** Exibe o desvio padrão, o erro padrão e o intervalo de confiança de 95% para o modelo de efeitos fixos, e o erro padrão, o intervalo de confiança de 95% e a estimativa da variância entre componentes para o modelo de efeitos aleatórios.
- **Homogeneidade de teste de variância.** Calcula a estatística de Levene para testar a igualdade de variâncias de grupos. Esse teste não depende da suposição de normalidade.

- **Brown-Forsythe.** Calcula a estatística de Brown-Forsythe para testar a igualdade das médias de grupo. Essa estatística é preferível sobre a estatística de *F* quando a suposição de variâncias iguais não é mantida.
- **Welch.** Calcula a estatística de Welch para testar a igualdade das médias de grupo. Essa estatística é preferível sobre a estatística de *F* quando a suposição de variâncias iguais não é mantida.

Gráfico de médias. Exibe um gráfico que representa as médias de subgrupo (as médias para cada grupo definidas por valores da variável de fator).

Valores Omissos. Controla o tratamento de valores omissos.

- **Excluir análise de casos por análise.** Um caso com um valor omisso para a variável dependente ou de fator de uma determinada análise não é utilizado nessa análise. Além disso, um caso fora do intervalo especificado para a variável de fator não é utilizado.
- **Excluir listwise de casos.** Casos com valores omissos para a variável de fator ou para qualquer variável dependente incluída na lista de variáveis dependentes na caixa de diálogo principal são excluídos de todas as análises. Se diversas variáveis dependentes não tiverem sido especificadas, isso não terá efeito.

Recursos Adicionais do Comando ONEWAY

O idioma da sintaxe de comando também permite:

- Obter estatísticas de efeitos fixo e aleatório. Desvio padrão, erro padrão da média e intervalos de confiança de 95% para o modelo de efeitos fixos. Erro padrão, intervalos de confiança de 95% e a estimativa de variância entre componentes para modelo de efeitos aleatórios (utilizando `STATISTICS=EFFECTS`).
- Especificar níveis alfa para a diferença menos significativa, Bonferroni, Duncan, e testes de comparação múltipla de Scheffé (com o subcomando `RANGES`).
- Gravar uma matriz de médias, desvios padrão e frequências, ou ler uma matriz de médias, frequências, variâncias agrupadas e graus de liberdade para as variâncias agrupadas. Essas matrizes podem ser utilizadas no lugar de dados brutos para obter uma análise de variância unidirecional (com o subcomando `MATRIX`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 11. Análise Univariate GLM

O procedimento do GLM Univariate fornece análise de regressão e a análise de variância para uma variável dependente por um ou mais fatores e/ou variáveis. As variáveis de fator dividem a população em grupos. Utilizando este procedimento Modelo Linear Geral, é possível testar hipóteses nulas sobre os efeitos de outras variáveis nas médias de vários agrupamentos de uma variável dependente única. É possível investigar as interações entre os fatores e também os efeitos de fatores individuais, alguns dos quais podendo ser aleatórios. Além disso, os efeitos de covariáveis e de interações de covariáveis com fatores podem ser incluídos. Para análise de regressão, as variáveis independentes (preditoras) são especificadas como covariáveis.

Modelos balanceados e não balanceados podem ser testados. Um design será balanceado se cada célula no modelo contiver o mesmo número de casos. Além de testar hipóteses, o GLM Univariate produz estimativas paramétricas.

Contrastes a priori normalmente utilizados estão disponíveis para executar teste de hipótese. Além disso, após um teste *F* geral ter mostrado significância, será possível utilizar testes post hoc para avaliar diferenças entre médias específicas. As médias marginais estimadas fornecem estimativas de valores médios preditos para as células no modelo, e os gráficos de perfil (gráficos de interação) dessas médias permitem visualizar facilmente alguns dos relacionamentos.

Valores residuais, valores preditos, a distância de Cook e valores de ponto de alavanca podem ser salvos como novas variáveis no arquivo de dados para verificação de suposições.

A Ponderação de WLS permite especificar uma variável utilizada para atribuir às observações ponderações diferentes para análise de quadrados mínimos ponderados (WLS), talvez para compensar uma precisão diferente de medição.

Exemplo. Os dados são reunidos para corredores individuais na maratona de Chicago por vários anos. O horário em que cada corredor termina a prova é a variável dependente. Outros fatores incluem o clima (frio, ameno ou calor), o número de meses de treinamento, número de maratonas anteriores e sexo. A idade é considerada uma covariável. Talvez você ache que sexo é um efeito significativo e que a interação de sexo com o clima é significativa.

Métodos. Somas dos quadrados do Tipo I, Tipo II, Tipo III e Tipo IV podem ser utilizadas para avaliar diferentes hipóteses. O Tipo III é o padrão.

Estatísticas. Testes de intervalo e comparações múltiplas autoteste inicial posteriori: diferença menos significativa, Bonferroni, Sidak, Scheffé, *F* múltiplo de Ryan-Einot-Gabriel-Welsch e amplitude múltipla de Ryan-Einot-Gabriel-Welsch, Aluno-Newman-Keuls, diferença honestamente significativa de Tukey, *b* de Tukey, Duncan, GT2 de Hochberg, Gabriel, teste *t* de Waller-Duncan, Dunnett (unilateral e bilateral), T2 de Tamhane, T3 de Dunnett, Games-Howell e *C* de Dunnet. Estatística descritiva: médias observadas, desvios padrão e contagens de todas as variáveis dependentes em todas as células. O teste de homogeneidade de variâncias de Levene.

Gráficos. Dispersão versus nível, residual e perfil (interação).

Considerações de Dados de GLM Univariate

Dados. A variável dependente é quantitativa. Os fatores são categóricos. Eles podem ter valores numéricos ou valores de sequência de caracteres de até oito caracteres. Covariáveis são variáveis quantitativas que estão relacionadas à variável dependente.

Suposições. Os dados são uma amostra aleatória de uma população normal; na população, todas as variâncias de célula são as mesmas. A análise de variância é robusta para partidas da normalidade, embora os dados devam ser simétricos. Para verificar suposições, é possível utilizar gráficos de testes de homogeneidade de variâncias e de dispersão versus nível. Também é possível examinar residuais e gráficos de resíduos.

Para Obter Tabelas GLM Univariate

1. Nos menus, escolha:

   Analisar > **Modelo Linear Geral** > **Univariada...**
2. Selecione uma variável dependente.
3. Selecione variáveis para Fator(es) Fixo(s), Fator(es) Aleatório(s) e Covariável(is), conforme apropriado para seus dados.
4. Opcionalmente, é possível utilizar Ponderação de WLS para especificar uma variável de ponderação para análise de quadrados mínimos ponderados. Se o valor da variável de ponderação for zero, negativo ou omisso, o caso será excluído da análise. Uma variável já utilizada no modelo não pode ser utilizada como uma variável de ponderação.

Modelo de GLM

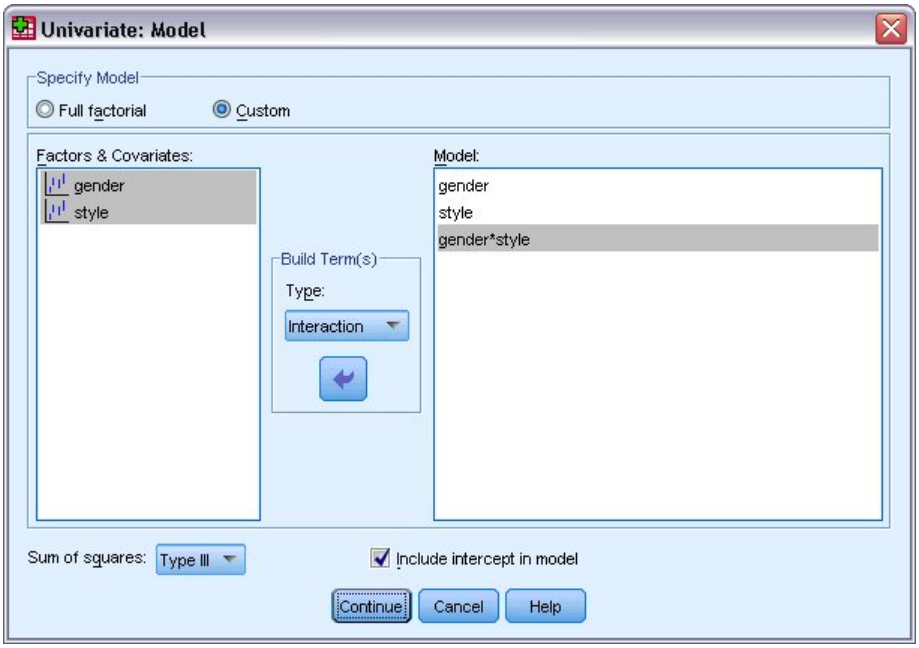

Figura 1. Caixa de diálogo Modelo univariado

Especificar modelo. Um modelo fatorial completo contém todos os principais efeitos do fator, todos os principais efeitos covariáveis e todas as interações fator por fator. Ele não contém interações covariáveis. Selecione **Customizado** para especificar apenas um subconjunto de interações ou para especificar interações fator por covariável. Deve-se indicar todos os termos a serem incluídos no modelo.

Fatores e covariáveis. Os fatores e covariáveis são listados.

Modelo. O modelo depende da natureza de seus dados. Depois de selecionar **Customizado**, é possível selecionar os principais efeitos e interações que são de interesse em sua análise.

Soma dos quadrados. O método de calcular as somas dos quadrados. Para modelos balanceados ou não balanceados sem células omissas, o método da soma dos quadrados do Tipo III é mais comumente usado.

Incluir intercepto no modelo. O intercepto geralmente é incluído no modelo. Se for possível presumir que os dados percorrerão a origem, será possível excluir o intercepto.

Termos de compilação

Para os fatores e covariáveis selecionados:

Interação. Cria o termo de interação de nível mais alto de todas as variáveis selecionadas. Esse é o padrão.

Efeitos principais. Cria um termo dos principais efeitos para cada variável selecionada.

Todos os 2 fatores. Cria todas as interações de dois fatores possíveis das variáveis selecionadas.

Todas 3 fatores. Cria todas as interações de três fatores das variáveis selecionadas.

Todos os 4 fatores. Cria todas as interações de quatro fatores possíveis das variáveis selecionadas.

Todos os 5 fatores. Cria todas as interações de cinco fatores possíveis das variáveis selecionadas.

Soma dos Quadrados

Para o modelo, é possível escolher um tipo de somas de quadrados. O tipo III é o mais comumente usado e é o padrão.

Tipo I. Esse método também é conhecido como decomposição hierárquica do método da soma dos quadrados. Cada termo é ajustado para apenas o termo que o precede no modelo. As somas dos quadrados do tipo I são comumente utilizados para:

- Um modelo ANOVA balanceado, no qual quaisquer efeitos principais serão especificados antes de quaisquer efeitos de interação de primeira ordem, quaisquer efeitos de interação de primeira ordem são especificados antes de quaisquer efeitos de interação de segunda ordem, e assim por diante.
- Um modelo de regressão polinomial no qual quaisquer termos de ordem inferior são especificados antes de quaisquer termos de ordem superior.
- Um modelo puramente aninhado no qual o efeito primeiro especificado é aninhado dentro do segundo efeito especificado, o efeito segundo especificado é aninhado dentro do terceiro, e assim por diante. (Esta forma de aninhamento pode ser especificada usando a sintaxe.)

Tipo II. Este método calcula a soma dos quadrados de um efeito no modelo ajustado para todos os outros efeitos "apropriados". Um efeito apropriado é aquele que corresponde a todos os efeitos que não contêm o efeito que está sendo examinado. O método da soma dos quadrados do Tipo II é comumente utilizado para:

- Um modelo ANOVA balanceado.
- Qualquer modelo que possui apenas efeitos de fator principal.
- Qualquer modelo de regressão.
- Um design puramente aninhado. (Esta forma de aninhamento pode ser especificada usando a sintaxe.)

Tipo III. O padrão. Este método calcula a soma dos quadrados de um efeito no design como somas de quadrados, ajustado para quaisquer outros efeitos que não contêm o efeito, e ortogonal para quaisquer efeitos (se houver) que contêm o efeito. A soma dos quadrados do Tipo III possui uma vantagem principal nas quais eles são invariáveis com relação às frequências de célula contanto que o formato geral de estimabilidade permaneça constante. Portanto, esse tipo de somas de quadrados é geralmente

considerado útil para um modelo não balanceado sem células omissas. Em um design fatorial sem células omissas, este método é equivalente à técnica Yates' weighted-squares-of-means. O método da soma dos quadrados do Tipo III é comumente utilizado para:

- Todos os modelos listados em Tipo I e Tipo II.
- Qualquer modelo balanceado ou não balanceado sem células vazias.

Tipo IV. Este método é projetado para uma situação em que houver células omissas. Para qualquer efeito *F* no design, se *F* não está contido em nenhum outro efeito, então Tipo IV = Tipo III = Tipo II. Quando *F* está contido em outros efeitos, o Tipo IV distribui os contrastes que estão sendo feitos entre os parâmetros em *F* para todos os efeitos de nível superior equitativamente. O método da soma dos quadrados do Tipo IV é comumente utilizado para:

- Todos os modelos listados em Tipo I e Tipo II.
- Qualquer modelo balanceado ou não balanceado com células vazias.

Contrastes GLM

Contrastes são utilizados para testar as diferenças entre os níveis de um fator. É possível especificar um contraste para cada fator no modelo (em um modelo de medidas repetidas, para cada fator entre assuntos). Os contrastes representam as combinações lineares dos parâmetros.

GLM Univariate. Os testes de hipótese baseiam-se na hipótese nula **LB** = 0, em que **L** é a matriz de coeficientes de contraste e **B** é o vetor paramétrica. Quando um contraste é especificado, uma matriz **L** é criada. As colunas da matriz **L** correspondem ao fator que corresponde ao contraste. As colunas restantes são ajustadas para que a matriz **L** seja estimável.

A saída inclui uma estatística *F* para cada conjunto de contrastes. Também exibidos para as diferenças de contrastes estão os intervalos de confiança simultâneos do tipo Bonferroni com base na distribuição *t* de Student.

Contrastes Disponíveis

Os contrastes disponíveis são desvio, simples, diferença, Helmert, repetido e polinomial. Para contrastes de desvio e contrastes simples, é possível escolher se a categoria de referência será a primeira ou a última categoria.

Tipos de Contraste

Desvio. Compara a média de cada nível (exceto uma categoria de referência) com a média de todos os níveis (média global). Os níveis do fator podem estar em qualquer ordem.

Simples. Compara a média de cada nível com a média de um nível especificado. Esse tipo de contraste é útil quando há um grupo de controle. É possível escolher a primeira ou a última categoria como a referência.

Diferença. Compara a média de cada nível (exceto do primeiro) com a média dos níveis anteriores. (Às vezes chamada de contrastes de Helmert reversos.)

Helmert. Compara a média de cada nível do fator (exceto do último) com a média dos níveis subsequentes.

Repetido. Compara a média de cada nível (exceto do último) com a média dos níveis subsequentes.

Polinomial. Compara o efeito linear, o efeito quadrático, o efeito cúbico, e assim por diante. O primeiro grau de liberdade contém o efeito linear em todas as categorias, o segundo grau de liberdade, o efeito quadrático, e assim por diante. Esses contrastes são frequentemente utilizados para estimar as tendências polinomiais.

Gráficos de perfil GLM

Os gráficos de perfil (gráficos de interação) são úteis para comparar médias marginais em seu modelo. Um gráfico de perfil é um gráfico de linha no qual cada ponto indica a média marginal estimada de uma variável dependente (ajustada para quaisquer covariáveis) em um nível de um fator. Os níveis de um segundo fator podem ser usados para fazer linhas separadas. Cada nível em um terceiro fator pode ser usado para criar um gráfico separado. Todos os fatores fixos e aleatórios, se houver, estão disponíveis para gráficos. Para análises multivariadas, gráficos de perfil são criados para cada variável dependente. Em uma análise de medidas repetidas, ambos os fatores entre assuntos e dentre-sujeitos podem ser utilizados em gráficos de perfil. As Medidas Multivariadas de GLM e Repetidas de GLM estão disponíveis somente se você tiver a opção Estatísticas Avançadas instalada.

Um gráfico de perfil de um fator mostra se as médias marginais estimadas estão aumentando ou diminuindo entre níveis. Para dois ou mais fatores, linhas paralelas indicam que não existe interação entre fatores, o que significa que é possível investigar os níveis de apenas um fator. Linhas não paralelas indicam uma interação.

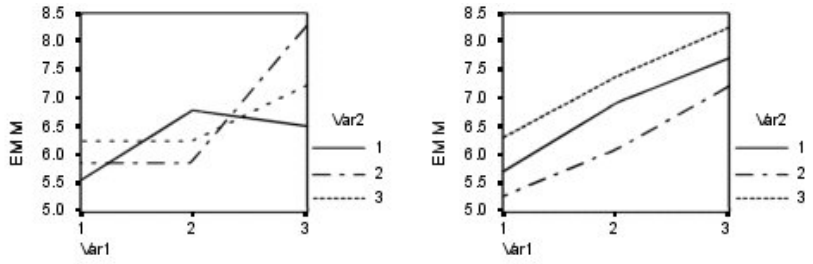

Figura 2. Gráfico não paralelo (esquerda) e gráfico paralelo (direita)

Depois de um gráfico ser especificado selecionando os fatores para o eixo horizontal e, opcionalmente, os fatores para linhas separadas e gráficos separados, o gráfico deve ser incluído na lista Gráficos.

Opções de GLM

Estatísticas de opcionais estão disponíveis a partir dessa caixa de diálogo. As estatísticas são calculadas utilizando um modelo de efeitos fixos.

Médias Marginais Estimadas. Selecione os fatores e interações para os quais deseja estimar as médias marginais da população nas células. Essas médias são ajustadas para as covariáveis, se houver.

- **Compare os efeitos principais.** Fornece comparações entre pares não corrigidas entre as médias marginais estimadas para qualquer efeito principal no modelo, para fatores entre e dentre-sujeitos. Esse item estará disponível somente se os efeitos principais forem selecionados na lista Médias de Exibição.
- **Ajustamento de intervalo de confiança.** Selecione o ajustamento de diferença menos significativa (LSD), Bonferroni ou Sidak ou para os intervalos de confiança e significância. Este item estará disponível apenas se **Comparar os efeitos principais** for selecionada.

Exibição. Selecione **Estatísticas descritivas** para produzir médias observadas, desvios padrão e contagens para todas as variáveis dependentes em todas as células. **Estimativas de tamanho de efeito** fornece um valor eta quadrado parcial para cada efeito e cada estimativa paramétrica. A estatística eta quadrado descreve a proporção da variabilidade total atribuível a um fator. Selecione **Potência observada** para obter a potência do teste quando a hipótese alternativa é configurada com base no valor observado.

Selecione **Estimativas paramétrica** para produzir as estimativas paramétrica, os erros padrão, testes *t*, intervalos de confiança e a potência observada para cada teste. Selecione **Matriz de coeficientes de contraste** para obter a matriz **L**.

Os **Testes de Homogeneidade** produzem o teste de Levene da homogeneidade da variância para cada variável dependente em todas as combinações de nível somente para os fatores entre assuntos. As opções de gráficos de dispersão versus nível e de resíduos são úteis para verificar suposições sobre os dados. Esse item fica desativado se não houver fatores. Selecione **Gráfico de resíduo** para produzir um gráfico residual observado por predito por padronizado para cada variável dependente. Esses gráficos são úteis para investigar a suposição de variância igual. Selecione **Falta de ajuste** para verificar se o relacionamento entre a variável dependente e as variáveis independentes pode ser descrito adequadamente pelo modelo. A **função estimável geral** permite construir testes de hipótese customizados com base na função estimável geral. As linhas em qualquer matriz de coeficientes de contraste são combinações lineares da função estimável geral.

Nível de significância. Talvez você queira ajustar o nível de significância utilizado em testes post hoc e o nível de confiança utilizado para construir intervalos de confiança. O valor especificado é também usado para calcular a potência observada para o teste. Ao especificar um nível de significância, o nível associado dos intervalos de confiança é exibido na caixa de diálogo.

Recursos Adicionais do Comando UNIANOVA

O idioma da sintaxe de comando também permite:

- Especificar efeitos aninhados no design (utilizando o subcomando `DESIGN`).
- Especificar testes de efeitos versus uma combinação linear de efeitos ou um valor (utilizando o subcomando `TEST`).
- Especificar diversos contrastes (utilizando o subcomando `CONTRAST`).
- Incluir valores omissos de usuário (utilizando o subcomando `MISSING`).
- Especificar critérios do ESP (utilizando o subcomando `CRITERIA`).
- Construir uma matriz **L**, uma matriz **M** ou uma matriz **K** customizada (usando os subcomandos `LMATRIX`, `MMATRIX` e `KMATRIX`).
- Para desvio ou contrastes simples, especifique uma categoria de referência intermediária (utilizando o subcomando `CONTRAST`).
- Especificar métricas para contrastes polinomiais (utilizando o subcomando `CONTRAST`).
- Especificar termos de erro para comparações post hoc (utilizando o subcomando `POSTHOC`).
- Calcular médias marginais estimadas para qualquer fator ou interação entre fatores entre os fatores na lista de fatores (utilizando o subcomando `EMMEANS`).
- Especificar nomes para variáveis temporárias (utilizando o subcomando `SAVE`).
- Construir um arquivo de dados de matriz de correlações (utilizando o subcomando `OUTFILE`).
- Construir um arquivo de dados de matriz que contenha estatísticas da tabela ANOVA entre assuntos (utilizando o subcomando `OUTFILE`).
- Salvar a matriz de design em um novo arquivo de dados (utilizando o subcomando `OUTFILE`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Comparações Posteriori do GLM

Testes de comparação múltipla post hoc. Após determinar que diferenças existem entre as médias, testes de intervalo post hoc e comparações múltiplas entre pares podem determinar quais médias diferem. As comparações são feitas em valores sem ajuste. Esses testes são usados apenas para fatores entre sujeitos. Em Medidas repetidas de GLM, esses testes não estarão disponíveis se não houver nenhum fator entre sujeitos, e os testes de múltiplas comparações post hoc são executados para a média nos níveis dos fatores dentro de sujeitos. Para GLM Multivariado, os testes post hoc são executados para cada variável

dependente separadamente. As Medidas Multivariadas de GLM e Repetidas de GLM estão disponíveis somente se você tiver a opção Estatísticas Avançadas instalada.

Os testes da diferença significativa honesta de Bonferroni e Tukey geralmente são usados por testes de comparação múltipla. O **teste de Bonferroni**, baseado na estatística *t* do aluno, ajusta o nível de significância observado para o fato de que comparações múltiplas são feitas. O **teste t de Sidak** também ajusta o nível de significância e fornece limites mais apertados que o teste de Bonferroni. O **teste da diferença significativa honesta de Tukey** usa o intervalo estatístico estudentizado para fazer todas as comparações entre pares entre os grupos e define a taxa de erro entre experimentos na taxa de erros da coleção de todas as comparações entre pares. Ao testar um grande número de pares de médias, o teste da diferença significativa honesta de Tukey é mais poderoso que o teste de Bonferroni. Para um pequeno número de pares, Bonferroni é mais poderoso.

GT2 de Hochberg é semelhante ao teste da diferença significativa honesta de Tukey, mas o módulo máximo estudentizado é usado. Geralmente, o teste de Tukey é mais poderoso. O **teste de comparação entre pares de Gabriel** também usa o módulo máximo estudentizado e geralmente é mais poderoso que o GT2 de Hochberg quando os tamanhos de células são desiguais. O teste de Gabriel pode ser liberal quando os tamanhos de células variam grandiosamente.

Teste t de comparação múltipla entre pares de Dunnett compara um conjunto de tratamentos com relação a uma média de controle única. A última categoria é a categoria de controle padrão. Como alternativa, é possível escolher a primeira categoria. Também é possível escolher um teste de dois lados ou unilateral. Para testar se a média em qualquer nível (exceto a categoria de controle) do fator não é igual à da categoria de controle, use um teste de dois lados. Para testar se a média em qualquer nível do fator é menor do que a da categoria de controle, selecione **< Controle**. Da mesma forma, para testar se a média em qualquer nível do fator é maior que a da categoria de controle, selecione **> Controle**.

Ryan, Einot, Gabriel e Welsch (R-E-G-W) desenvolveram dois testes de intervalo de redução múltipla. Os procedimentos de redução múltipla primeiro testam se todas as médias são iguais. Se todas as médias não forem iguais, os subconjuntos de médias serão testadas para igualdade. **R-E-G-W F** é baseado em um teste *F*, e **R-E-G-W Q** é baseado no intervalo estudentizado. Esses testes são mais poderosos do que o teste de amplitude múltipla de Duncan e o Student-Newman-Keuls (que também são procedimentos de redução múltipla), mas eles não são recomendados para os tamanhos de células desiguais.

Quando as variâncias forem desiguais, use **T2 de Tamhane** (teste de comparação entre pares conservador com base em um teste *t*), **T3 de Dunnett** (teste de comparação entre pares com base no Módulo máximo estudentizado), **teste de comparação entre pares Games-Howell** (às vezes, liberals) ou o **C de Dunnett** (teste de comparação entre pares com base no intervalo estudentizado). Observe que esses testes não são válidos e não serão produzidos se houver vários fatores no modelo.

O **teste de amplitude múltipla de Duncan**, Student-Newman-Keuls (**S-N-K**) e o **b de Tukey** são testes de intervalo que classificam médias de grupo e calculam um valor de intervalo. Esses testes não são usados tão frequentemente quanto os testes discutidos anteriormente.

O **teste t de Waller-Duncan** usa uma abordagem bayesiana. Esse teste de intervalo usa a média harmônica do tamanho da amostra quando os tamanhos da amostra são desiguais.

O nível de significância do teste **Scheffé** foi projetado para permitir que todas as combinações lineares possíveis das médias de grupo fossem testadas, não apenas as comparações entre pares disponíveis neste recurso. O resultado é que o teste de Scheffé é geralmente mais conservador que outros testes, o que significa que uma diferença maior entre as médias é necessária para obter significância.

O teste de múltipla comparação entre pares da diferença menos significativa (**LSD**) é equivalente a testes *t* individuais múltiplos entre todos os pares de grupos. A desvantagem deste teste é que nenhuma tentativa será feita para ajustar o nível de significância observado para comparações múltiplas.

Testes exibidos. As comparações entre pares são fornecidas para LSD, Sidak, Bonferroni, Games-Howell, T2 e T3 de Tamhane, Cde Dunnett e T3 de Dunnett. Os subconjuntos homogêneos para testes de intervalo são fornecidas para S-N-K, *b* de Tukey, Duncan, R-E-G-W *F*, R-E-G-W *Q* e Waller. O teste da diferença significativa honesta de Tukey, o GT2 de Hochberg, o teste de Gabriel e o teste de Scheffé são testes de comparação múltipla e testes de intervalo.

Opções de GLM

Estatísticas de opcionais estão disponíveis a partir dessa caixa de diálogo. As estatísticas são calculadas utilizando um modelo de efeitos fixos.

Médias Marginais Estimadas. Selecione os fatores e interações para os quais deseja estimar as médias marginais da população nas células. Essas médias são ajustadas para as covariáveis, se houver.

- **Compare os efeitos principais.** Fornece comparações entre pares não corrigidas entre as médias marginais estimadas para qualquer efeito principal no modelo, para fatores entre e dentre-sujeitos. Esse item estará disponível somente se os efeitos principais forem selecionados na lista Médias de Exibição.
- **Ajustamento de intervalo de confiança.** Selecione o ajustamento de diferença menos significativa (LSD), Bonferroni ou Sidak ou para os intervalos de confiança e significância. Este item estará disponível apenas se **Comparar os efeitos principais** for selecionada.

Exibição. Selecione **Estatísticas descritivas** para produzir médias observadas, desvios padrão e contagens para todas as variáveis dependentes em todas as células. **Estimativas de tamanho de efeito** fornece um valor eta quadrado parcial para cada efeito e cada estimativa paramétrica. A estatística eta quadrado descreve a proporção da variabilidade total atribuível a um fator. Selecione **Potência observada** para obter a potência do teste quando a hipótese alternativa é configurada com base no valor observado. Selecione **Estimativas paramétrica** para produzir as estimativas paramétrica, os erros padrão, testes *t*, intervalos de confiança e a potência observada para cada teste. Selecione **Matriz de coeficientes de contraste** para obter a matriz **L**.

Os **Testes de Homogeneidade** produzem o teste de Levene da homogeneidade da variância para cada variável dependente em todas as combinações de nível somente para os fatores entre assuntos. As opções de gráficos de dispersão versus nível e de resíduos são úteis para verificar suposições sobre os dados. Esse item fica desativado se não houver fatores. Selecione **Gráfico de resíduo** para produzir um gráfico residual observado por predito por padronizado para cada variável dependente. Esses gráficos são úteis para investigar a suposição de variância igual. Selecione **Falta de ajuste** para verificar se o relacionamento entre a variável dependente e as variáveis independentes pode ser descrito adequadamente pelo modelo. A **função estimável geral** permite construir testes de hipótese customizados com base na função estimável geral. As linhas em qualquer matriz de coeficientes de contraste são combinações lineares da função estimável geral.

Nível de significância. Talvez você queira ajustar o nível de significância utilizado em testes post hoc e o nível de confiança utilizado para construir intervalos de confiança. O valor especificado é também usado para calcular a potência observada para o teste. Ao especificar um nível de significância, o nível associado dos intervalos de confiança é exibido na caixa de diálogo.

Recursos Adicionais do Comando UNIANOVA

O idioma da sintaxe de comando também permite:

- Especificar efeitos aninhados no design (utilizando o subcomando `DESIGN`).
- Especificar testes de efeitos versus uma combinação linear de efeitos ou um valor (utilizando o subcomando `TEST`).
- Especificar diversos contrastes (utilizando o subcomando `CONTRAST`).
- Incluir valores omissos de usuário (utilizando o subcomando `MISSING`).
- Especificar critérios do ESP (utilizando o subcomando `CRITERIA`).

- Construir uma matriz **L**, uma matriz **M** ou uma matriz **K** customizada (usando os subcomandos LMATRIX, MMATRIX e KMATRIX).
- Para desvio ou contrastes simples, especifique uma categoria de referência intermediária (utilizando o subcomando CONTRAST).
- Especificar métricas para contrastes polinomiais (utilizando o subcomando CONTRAST).
- Especificar termos de erro para comparações post hoc (utilizando o subcomando POSTHOC).
- Calcular médias marginais estimadas para qualquer fator ou interação entre fatores entre os fatores na lista de fatores (utilizando o subcomando EMMEANS).
- Especificar nomes para variáveis temporárias (utilizando o subcomando SAVE).
- Construir um arquivo de dados de matriz de correlações (utilizando o subcomando OUTFILE).
- Construir um arquivo de dados de matriz que contenha estatísticas da tabela ANOVA entre assuntos (utilizando o subcomando OUTFILE).
- Salvar a matriz de design em um novo arquivo de dados (utilizando o subcomando OUTFILE).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Salvamento de GLM

É possível salvar valores preditos pelo modelo, resíduos e medidas relacionadas como novas variáveis no Editor de Dados. Muitas dessas variáveis podem ser usadas para examinar suposições sobre os dados. Para salvar os valores para usar em outra sessão do IBM SPSS Statistics, deve-se salvar o arquivo de dados atual

Valores Preditos. Os valores que o modelo prediz para cada caso.

- *Não padronizado*. O valor que o modelo prediz para a variável dependente.
- *Ponderado*. Valores preditos não padronizados ponderados. Disponível apenas se uma variável WLS foi selecionada anteriormente.
- *Erro padrão*. Uma estimativa do desvio padrão do valor médio da variável dependente para os casos que tiverem os mesmos valores das variáveis independentes.

Diagnósticos. Medidas para identificar os casos com combinações de valores incomuns para as variáveis independentes e os casos que podem ter um grande impacto no modelo.

- *Distância de cook*. Uma medida do quanto os resíduos de todos os casos seriam alterados se um caso específico fosse excluído do cálculo dos coeficientes de regressão. Um D de Cook grande indica que excluir um caso do cálculo das estatísticas de regressão altera os coeficientes substancialmente.
- *REMOVE*. Valores de ponto de alavanca não centralizados. A influência relativa de cada observação sobre o ajuste do modelo.

Residuais. Um resíduo não padronizado é o valor real da variável dependente menos o valor predito pelo modelo. Resíduos padronizados, estudantizados e excluídos também estão disponíveis. Se uma variável WLS foi escolhida, resíduos não padronizados ponderados estarão disponíveis.

- *Não padronizado*. A diferença entre um valor observado e o valor predito pelo modelo
- *Ponderado*. Resíduos não padronizados ponderados. Disponível apenas se uma variável WLS foi selecionada anteriormente.
- *Padronizado*. O resíduo dividido por uma estimativa do seu desvio padrão. Resíduos padronizados, também conhecidos como resíduos de Pearson, possuem uma média de 0 e um desvio padrão de 1.
- *Estudentizado*. O resíduo dividido por uma estimativa do seu desvio padrão que varia de caso para caso, dependendo da distância dos valores de cada caso nas variáveis independentes das médias das variáveis independentes.
- *Excluído*. O resíduo para um caso quando esse caso é excluído do cálculo dos coeficientes de regressão. Ele é a diferença entre o valor da variável dependente e o valor predito ajustado.

Estatísticas de coeficiente. Grava uma matriz de variância-covariância das estimativas do parâmetro no modelo em um novo conjunto de dados na sessão atual ou em um arquivo de dados externo do IBM SPSS Statistics. Além disso, para cada variável dependente, haverá uma linha de estimativas paramétrica, uma linha de valores de significância para as estatísticas *t* correspondentes às estimativas paramétrica e uma linha de graus de liberdade de resíduos. Para um modelo multivariado, há linhas semelhantes para cada variável dependente. É possível usar este arquivo de matriz em outros procedimentos que leem arquivos de matriz.

Opções de GLM

Estatísticas de opcionais estão disponíveis a partir dessa caixa de diálogo. As estatísticas são calculadas utilizando um modelo de efeitos fixos.

Médias Marginais Estimadas. Selecione os fatores e interações para os quais deseja estimar as médias marginais da população nas células. Essas médias são ajustadas para as covariáveis, se houver.

- **Compare os efeitos principais.** Fornece comparações entre pares não corrigidas entre as médias marginais estimadas para qualquer efeito principal no modelo, para fatores entre e dentre-sujeitos. Esse item estará disponível somente se os efeitos principais forem selecionados na lista Médias de Exibição.
- **Ajustamento de intervalo de confiança.** Selecione o ajustamento de diferença menos significativa (LSD), Bonferroni ou Sidak ou para os intervalos de confiança e significância. Este item estará disponível apenas se **Comparar os efeitos principais** for selecionada.

Exibição. Selecione **Estatísticas descritivas** para produzir médias observadas, desvios padrão e contagens para todas as variáveis dependentes em todas as células. **Estimativas de tamanho de efeito** fornece um valor eta quadrado parcial para cada efeito e cada estimativa paramétrica. A estatística eta quadrado descreve a proporção da variabilidade total atribuível a um fator. Selecione **Potência observada** para obter a potência do teste quando a hipótese alternativa é configurada com base no valor observado. Selecione **Estimativas paramétrica** para produzir as estimativas paramétrica, os erros padrão, testes *t*, intervalos de confiança e a potência observada para cada teste. Selecione **Matriz de coeficientes de contraste** para obter a matriz **L**.

Os **Testes de Homogeneidade** produzem o teste de Levene da homogeneidade da variância para cada variável dependente em todas as combinações de nível somente para os fatores entre assuntos. As opções de gráficos de dispersão versus nível e de resíduos são úteis para verificar suposições sobre os dados. Esse item fica desativado se não houver fatores. Selecione **Gráfico de resíduo** para produzir um gráfico residual observado por predito por padronizado para cada variável dependente. Esses gráficos são úteis para investigar a suposição de variância igual. Selecione **Falta de ajuste** para verificar se o relacionamento entre a variável dependente e as variáveis independentes pode ser descrito adequadamente pelo modelo. A **função estimável geral** permite construir testes de hipótese customizados com base na função estimável geral. As linhas em qualquer matriz de coeficientes de contraste são combinações lineares da função estimável geral.

Nível de significância. Talvez você queira ajustar o nível de significância utilizado em testes post hoc e o nível de confiança utilizado para construir intervalos de confiança. O valor especificado é também usado para calcular a potência observada para o teste. Ao especificar um nível de significância, o nível associado dos intervalos de confiança é exibido na caixa de diálogo.

Recursos Adicionais do Comando UNIANOVA

O idioma da sintaxe de comando também permite:

- Especificar efeitos aninhados no design (utilizando o subcomando `DESIGN`).
- Especificar testes de efeitos versus uma combinação linear de efeitos ou um valor (utilizando o subcomando `TEST`).
- Especificar diversos contrastes (utilizando o subcomando `CONTRAST`).

- Incluir valores omissos de usuário (utilizando o subcomando MISSING).
- Especificar critérios do ESP (utilizando o subcomando CRITERIA).
- Construir uma matriz **L**, uma matriz **M** ou uma matriz **K** customizada (usando os subcomandos LMATRIX, MMATRIX e KMATRIX).
- Para desvio ou contrastes simples, especifique uma categoria de referência intermediária (utilizando o subcomando CONTRAST).
- Especificar métricas para contrastes polinomiais (utilizando o subcomando CONTRAST).
- Especificar termos de erro para comparações post hoc (utilizando o subcomando POSTHOC).
- Calcular médias marginais estimadas para qualquer fator ou interação entre fatores entre os fatores na lista de fatores (utilizando o subcomando EMMEANS).
- Especificar nomes para variáveis temporárias (utilizando o subcomando SAVE).
- Construir um arquivo de dados de matriz de correlações (utilizando o subcomando OUTFILE).
- Construir um arquivo de dados de matriz que contenha estatísticas da tabela ANOVA entre assuntos (utilizando o subcomando OUTFILE).
- Salvar a matriz de design em um novo arquivo de dados (utilizando o subcomando OUTFILE).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 12. Correlações Bivariadas

O procedimento Correlações Bivariadas calcula o coeficiente de correlação de Pearson, o rô de Spearman e o tau-*b* de Kendall com seus níveis de significância. As correlações medem como as ordens de variáveis ou de ranqueamento estão relacionadas. Antes de calcular um coeficiente de correlação, verifique se seus dados possuem valores discrepantes (que podem causar resultados enganosos) e prova de um relacionamento linear. O coeficiente de correlação de Pearson é uma medida de associação linear. Duas variáveis podem estar perfeitamente relacionadas, mas se o relacionamento não for linear, o coeficiente de correlação de Pearson não será uma estatística apropriada para medir a associação delas.

Exemplo. O número de jogos vencidos por um time de basquete está correlacionado com o número médio de pontos marcados por jogo? Um gráfico de dispersão indica que há um relacionamento linear. A análise de dados das temporadas da NBA de 1994 e 1995 indica que o coeficiente de correlação de Pearson (0,581) é significativo no nível 0,01. Você pode suspeitar que quanto mais jogos foram vencidos por temporada, menos pontos os adversários marcaram. Essas variáveis são negativamente correlacionadas (-0,401), e a correlação é significativa no nível 0,05.

Estatísticas. Para cada variável: número de casos com valores não omissos, média e desvio padrão. Para cada par de variáveis: coeficiente de correlação de Pearson, rô de Spearman, tau-*b* de Kendall, produto cruzado de desvios e covariâncias.

Considerações de Dados de Correlações Bivariadas

Dados. Utilize variáveis quantitativas simétricas para o coeficiente de correlação de Pearson e variáveis quantitativas ou variáveis com categorias ordenadas por rô de Spearman e tau-*b* de Kendall.

Suposições. O coeficiente de correlação de Pearson supõe que cada par de variáveis é bivariado normal.

Para Obter Correlações Bivariadas

Nos menus, escolha:

Analisar > **Correlacionar** > **Bivariado...**

1. Selecione duas ou mais variáveis numéricas.

As seguintes opções também estão disponíveis:

- **Coeficientes de Correlação.** Para variáveis quantitativas normalmente distribuídas, escolha o coeficiente de correlação de **Pearson**. Se seus dados não estiverem normalmente distribuídos ou possuírem categorias ordenadas, escolha **tau-b de Kendall** ou **Spearman**, que mede a associação entre as ordens de ranqueamento. O valor do intervalo de coeficientes de correlação varia de -1 (um relacionamento negativo perfeito) e +1 (um relacionamento positivo perfeito). Um valor de 0 indica nenhum relacionamento linear. Ao interpretar os resultados, cuidado para não tirar nenhuma conclusão de causa e efeito devido a uma correlação significativa.
- **Teste de Significância.** É possível selecionar probabilidades bilaterais ou unilaterais. Se a direção da associação for conhecida com antecedência, selecione **Unilateral**. Caso contrário, selecione **Bilateral**.
- **Sinalizar correlações significativas.** Os coeficientes de correlação significativos no nível 0,05 são identificados com um asterisco único, e os significativos no nível 0,01 são identificados com dois asteriscos.

Opções de Correlações Bivariadas

Estatísticas. Para correlações de Pearson, é possível escolher uma ou as duas das seguintes opções:

- **Médias e desvios padrão.** Exibida para cada variável. O número de casos com valores não omissos também é mostrado. Os valores omissos são manipulados basicamente variável por variável, independentemente de sua configuração de valores omissos.
- **Desvios e covariâncias de produto vetorial.** Exibida para cada par de variáveis. O produto cruzado de desvios é igual à soma dos produtos das variáveis corrigidas pela média. Este é o numerador do coeficiente de correlação de Pearson. A covariância é uma medida não padronizada do relacionamento entre duas variáveis, igual ao desvio de produto vetorial dividido por *N*-1.

Valores omissos. É possível escolher uma das seguintes opções:

- **Excluir casos entre pares.** Casos com valores omissos para um ou ambos os elementos de um par de variáveis para um coeficiente de correlação são excluídos da análise. Como cada coeficiente baseia-se em todos os casos que tiverem códigos válidos nesse par específico de variáveis, o máximo de informações disponíveis é utilizado em cada cálculo. Isso pode resultar em um conjunto de coeficientes com base em um número variado de casos.
- **Excluir listwise dos casos.** Casos com valores omissos para qualquer variável são excluídos de todas as correlações.

Recursos Adicionais dos Comandos CORRELATIONS e NONPAR CORR

O idioma da sintaxe de comando também permite:

- Gravar uma matriz de correlações para correlações de Pearson que podem ser utilizadas no lugar de dados brutos para obter outras análises como análise fatorial (com o subcomando `MATRIX`).
- Obter correlações de cada variável em uma lista com cada variável em uma segunda lista (utilizando a palavra-chave `WITH` no subcomando `VARIABLES`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 13. Correlações parciais

O procedimento Correlações Parciais calcula coeficientes de correlação parcial que descrevem o relacionamento linear entre duas variáveis enquanto controla os efeitos de uma ou mais variáveis adicionais. As correlações são medidas de associação linear. Duas variáveis podem estar perfeitamente relacionadas, mas se o relacionamento não for linear, um coeficiente de correlação não será uma estatística apropriada para medir a associação delas.

Exemplo. Existe uma relação entre financiamento de assistência médica e taxas de doença? Embora talvez você espere que qualquer relacionamento desse tipo seja um relacionamento negativo, um estudo revela uma correlação *positiva* significativa: conforme o financiamento em assistência médica aumenta, taxas de doença parecem aumentar. No entanto, o controle da taxa de visitas nos provedores de assistência médica praticamente elimina a correlação positiva observada. O financiamento de assistência médica e as taxas de doença parecem estar positivamente relacionadas somente porque mais pessoas têm acesso à assistência médica quando o financiamento aumenta, levando a mais doenças diagnosticadas por médicos e hospitais.

Estatísticas. Para cada variável: número de casos com valores não omissos, média e desvio padrão. Matrizes de correlação parcial e de ordem zero, com graus de liberdade e níveis de significância.

Considerações de Dados de Correlações Parciais

Dados. Use variáveis quantitativas simétricas.

Suposições. O coeficiente de Correlações Parciais supõe que cada par de variáveis é bivariado normal.

Para Obter Correlações Parciais

1. Nos menus, escolha:

   Analisar > **Correlacionar** > **Parcial...**
2. Selecione duas ou mais variáveis numéricas para as quais as correlações parciais devem ser calculadas.
3. Selecione uma ou mais variáveis de controle numéricas.

As seguintes opções também estão disponíveis:

- **Teste de Significância.** É possível selecionar probabilidades bilaterais ou unilaterais. Se a direção da associação for conhecida com antecedência, selecione **Unilateral**. Caso contrário, selecione **Bilateral**.
- **Exibir o nível de significância real.** Por padrão, a probabilidade e os graus de liberdade são mostrados para cada coeficiente de correlação. Se cancelar a seleção desse item, os coeficientes significativos no nível 0,05 serão identificados com um asterisco único, os coeficientes significativos no nível 0,01 serão identificados com um asterisco duplo, e os graus de liberdade serão suprimidos. Essa configuração afeta matrizes de correlação parcial e de ordem zero.

Opções de Correlações Parciais

Estatísticas. É possível escolher um ou ambos os seguintes itens:

- **Médias e desvios padrão.** Exibida para cada variável. O número de casos com valores não omissos também é mostrado.
- **Correlações de ordem zero.** Uma matriz de correlações simples entre todas as variáveis, incluindo variáveis de controle, é exibida.

Valores omissos. É possível escolher uma das seguintes alternativas:

- **Excluir listwise dos casos.** Casos com valores omissos para qualquer variável, incluindo uma variável de controle, são excluídos de todos os cálculos.
- **Excluir casos entre pares.** Para o cálculo das correlações de ordem zero na qual as correlações parciais são baseadas, um caso que possuir valores omissos para ambos ou um par de variáveis não é utilizado. A exclusão dos pares utiliza o máximo de dados possível. No entanto, o número de casos pode diferir entre os coeficientes. Quando a exclusão dos pares estiver em vigor, os graus de liberdade de um determinado coeficiente parcial são baseados no menor número de casos utilizados no cálculo de qualquer uma das correlações de ordem zero.

Recursos Adicionais do Comando PARTIAL CORR

O idioma da sintaxe de comando também permite:

- Ler uma matriz de correlações de ordem zero ou gravar uma matriz de correlações parciais (com o subcomando `MATRIX`).
- Obter correlações parciais entre duas listas de variáveis (utilizando a palavra-chave `WITH` no subcomando `VARIABLES`).
- Obter análises múltiplas (com diversos subcomandos `VARIABLES`).
- Especificar valores do pedido para solicitação (por exemplo, correlações parciais de primeira e segunda ordem) quando houver duas variáveis de controle (com o subcomando `VARIABLES`).
- Suprimir coeficientes redundantes (com o subcomando `FORMAT`).
- Exibir uma matriz de correlações simples quando alguns coeficientes não podem ser calculados (com o subcomando `STATISTICS`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 14. Distâncias

Este procedimento calcula qualquer variedade ampla de estatísticas que medem similaridades ou dissimilaridades (distâncias), seja entre pares de variáveis ou entre pares de caso. Essas medidas de similaridade ou de distância podem então ser utilizadas com outros procedimentos, como análise fatorial, análise de cluster ou ajuste de escala multidimensional, para ajudar a analisar conjuntos de dados complexos.

Exemplo. É possível medir similaridades entre pares de automóveis com base em determinadas características, como tamanho do motor, Km/h e cavalo-vapor? Ao calcular similaridades entre automóveis, é possível ter uma ideia de quais automóveis são semelhantes e quais são diferentes um do outro. Para uma análise mais formal, é possível considerar a aplicação de uma análise de cluster hierárquica ou um ajuste de escala multidimensional às similaridades para explorar a estrutura subjacente.

Estatísticas. As medidas de dissimilaridade (distância) para dados do intervalo são distância euclidiana, distância euclidiana quadrática, Chebychev, bloco, Minkowski, ou customizadas; para dados de contagem, qui-quadrado ou fi-quadrado; para dados binários, a distância euclidiana, distância euclidiana quadrática, diferença de tamanho, diferença padrão, variância, forma ou Lance e Williams. As medidas de similaridade para dados do intervalo são correlação ou cosseno de Pearson; para dados binários, Russel e Rao, correspondência simples, Jaccard, divisão, 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* de Anderberg, *Y* Yule, *Q* de Yuli, Ochiai, Sokal e Sneath 5, correlação de fi de 4 pontos ou dispersão.

Para Obter Matrizes de Distância

1. Nos menus, escolha:

   Analisar > **Correlacionar** > **Distâncias...**
2. Selecione pelo menos uma variável numérica para calcular distâncias entre casos, ou selecione pelo menos duas variáveis numéricas para calcular distâncias entre as variáveis.
3. Selecione uma alternativa no grupo Calcular Distâncias para calcular proximidades entre os casos ou entre variáveis.

Medidas de Dissimilaridade de Distâncias

No grupo Medida, selecione a alternativa que corresponda ao seu tipo de dados (intervalo, contagem ou binário) e, em seguida, na lista suspensa, selecione uma das medidas que corresponda a esse tipo de dados. As medidas disponíveis, por tipo de dados, são:

- **Dados do intervalo.** Distância Euclidiana, distância euclidiana quadrática, Chebychev, bloco, Minkowski ou customizada.
- **Dados de contagem.** Medida qui-quadrado ou medida fi-quadrado.
- **Dados binários.** Distância euclidiana, distância euclidiana quadrática, diferença de tamanho, diferença de padrão, variância, forma ou Lance e Williams. (Insira valores para Presente e Ausente para especificar quais dois valores são significativos; Distâncias ignorará todos os outros valores).

O grupo Valores de Transformação permite padronizar valores de dados para quaisquer casos ou variáveis *antes* do cálculo de proximidades. Essas transformações não são aplicáveis aos dados binários. Os métodos de padronização disponíveis são escores *z*, intervalo -1 a 1, intervalo 0 a 1, magnitude máxima de 1, média de 1 ou desvio padrão de 1.

O grupo Medidas de Transformação permite transformar os valores gerados pela medida de distância. Eles são aplicados após a medida de distância ter sido calculada. As opções disponíveis são valores absolutos, sinal de mudança e reajuste de escala para o intervalo 0-1.

Medidas de Similaridade de Distâncias

No grupo Medida, selecione a alternativa que corresponda ao seu tipo de dados (intervalo ou binário) e, em seguida, na lista suspensa, selecione uma das medidas que corresponda a esse tipo de dados. As medidas disponíveis, por tipo de dados, são:

- **Dados do intervalo.** Correlação ou cosseno de Pearson.
- **Dados binários.** Russell e Rao, correspondência simples, 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* d Anderberg, *Y* de Yule, *Q* de Yule, Ochiai, Sokal e Sneath 5, correlação de fi de 4 pontos ou dispersão. (Insira valores para Presente e Ausente para especificar quais dois valores são significativos; Distâncias ignorará todos os outros valores).

O grupo Valores de Transformação permite padronizar valores de dados para quaisquer casos ou variáveis antes do cálculo de proximidades. Essas transformações não são aplicáveis aos dados binários. Os métodos de padronização disponíveis são escores *z*, intervalo -1 a 1, intervalo 0 a 1, magnitude máxima de 1, média de 1 e desvio padrão de 1.

O grupo Medidas de Transformação permite transformar os valores gerados pela medida de distância. Eles são aplicados após a medida de distância ter sido calculada. As opções disponíveis são valores absolutos, sinal de mudança e reajuste de escala para o intervalo 0-1.

Recursos Adicionais do Comando PROXIMITIES

O procedimento Distâncias utiliza a sintaxe de comando `PROXIMITIES`. O idioma da sintaxe de comando também permite:

- Especificar qualquer número inteiro como a potência para a medida de distância de Minkowski.
- Especificar quaisquer números inteiros como a potência e raiz para uma medida de distância customizada.

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 15. Modelos lineares

Os modelos lineares preveem uma variável resposta contínua com base em relacionamentos lineares entre a resposta e um ou mais preditores.

Modelos lineares são relativamente simples e fornecem uma fórmula matemática facilmente interpretada para escoragem. As propriedades desses modelos são bem entendidas e, normalmente, podem ser construídas muito rapidamente em comparação com outros tipos de modelo (como redes neurais ou árvores de decisão) no mesmo conjunto de dados.

Exemplo. Uma empresa de seguros com recursos limitados para investigar indenizações de seguro dos proprietários de imóveis quer construir um modelo para estimar os custos de indenizações. Implementando esse modelo em centros de serviço, os representantes podem inserir informações sobre indenização enquanto estiverem ao telefone com um cliente e obterem imediatamente o custo "esperado" da indenização com base em dados passados.

Requisitos de campo. Deve haver uma Resposta e pelo menos uma Entrada. Por padrão, campos com funções predefinidas de Ambos ou Nenhum não são usados. A resposta deve ser contínua (escala). Não há nenhuma restrição de nível de medição em preditores (entradas); campos (flag do nominal e ordinal) categóricos são usados como fatores no modelo e campos contínuos são usados como covariáveis

Nota: Se um campo categórico tiver mais de 1000 categorias, o procedimento não será executado e nenhum modelo será construído.

Para obter um modelo linear

Essa variável requer a opção de Base de Estatísticas.

Nos menus, escolha:

Analisar > **Regressão** > **Modelos Lineares Automáticos...**

1. Certifique-se de que exista pelo menos uma resposta e uma entrada.
2. Clique em **Opções de Criação** para especificar configurações de construção e de modelo opcionais.
3. Clique em **Opções de Modelo** para salvar escores no conjunto de dados ativo e exportar o modelo para um arquivo externo.
4. Clique em **Executar** para executar o procedimento e criar os objetos Modelo.

Objetivos

Qual é o seu objetivo principal? Selecione o objetivo apropriado.

- **Crie um modelo padrão.** O método constrói um modelo único para predizer a resposta usando os preditores. De um modo geral, modelos padrão são mais fáceis de interpretar e podem ser mais rápidos para pontuar do que combinações de conjuntos de dados impulsionadas, empacotadas ou grandes.
- **Aprimorar a precisão do modelo (boosting).** O método constrói um modelo de combinação usando boosting, que gera uma sequência de modelos para obter predições mais exatas. Os conjuntos podem demorar mais para construir e pontuar do que um modelo padrão.

  Boosting produz uma sucessão de "modelos de componente", cada um dos quais sendo construído no conjunto de dados inteiro. Antes de construir cada modelo de componente sucessivo, os registros são ponderados com base nos resíduos do modelo de componente anterior. Casos com grandes resíduos recebem ponderações de análise relativamente superiores, para que o próximo modelo de componente

se concentre também na predição desses registros. Juntos, esses modelos de componente formam um modelo de combinação. O modelo de combinação pontua novos registros usando uma regra de combinação; as regras disponíveis dependem do nível de medição da resposta.

- **Aprimorar estabilidade do modelo (bagging).** O método constrói um modelo de combinação usando bagging (agregação de bootstrap), que gera vários modelos para obter predições mais confiáveis. Os conjuntos podem demorar mais para construir e pontuar do que um modelo padrão.

  A agregação de bootstrap (bagging) produz réplicas do conjunto de dados de treinamento fazendo amostragem com substituição a partir do conjunto de dados original. Isso cria amostras bootstrap de tamanho igual ao conjunto de dados original. Em seguida, um "modelo de componente" é construído em cada réplica. Juntos, esses modelos de componente formam um modelo de combinação. O modelo de combinação pontua novos registros usando uma regra de combinação; as regras disponíveis dependem do nível de medição da resposta.
- **Criar um modelo para conjuntos de dados muito grandes (requer IBM SPSS Statistics Server).** O método constrói um modelo de combinação dividindo o conjunto de dados em blocos de dados separados. Escolha esta opção se o seu conjunto de dados for muito grande para construir qualquer um dos modelos acima, ou para construção de modelo incremental. Esta opção pode levar menos tempo para construir, mas pode demorar mais para pontuar do que um modelo padrão. Essa opção requer conectividade do IBM SPSS Statistics Server

Consulte “Combinações” na página 63 para obter as configurações relacionadas a boosting, bagging e conjuntos de dados muito grandes.

Básico

Preparar dados automaticamente. Essa opção permite que o procedimento transforme internamente a resposta e preditores para maximizar o poder preditivo do modelo; qualquer transformação é salva com o modelo e aplicada nos novos dados para escoragem. As versões originais de campos transformados são excluídas do modelo. Por padrão, a preparação de dado automático a seguir é executada.

- **Tratamento de Data e Hora.** Cada preditor de data é transformado em um novo preditor contínuo contendo o tempo decorrido desde uma data de referência (01-01-1970). Cada preditor de tempo é transformado em um novo preditor contínuo contendo o tempo decorrido desde um tempo de referência (00h00min00s).
- **Ajustar nível de medição.** Preditores contínuos com menos de 5 valores distintos são reformulados como preditores ordinais. Preditores ordinais com mais de 10 valores distintos são reformulados como preditores contínuos.
- **Tratamento de valor discrepante.** Valores de preditores contínuos que estão além de um valor de corte (3 desvios padrão a partir da média) são configurados para o valor de corte.
- **Tratamento de valor omisso.** Valores omissos de preditores nominais são substituídos pelo modo da partição de treinamento. Valores omissos de preditores ordinais são substituídos pela mediana da partição de treinamento. Valores omissos de preditores contínuos são substituídos pela média da partição de treinamento.
- **Mesclagem supervisionada.** Isso torna um modelo mais econômico reduzindo o número de campos a serem processados em associação com a resposta. Categorias semelhantes são identificadas com base no relacionamento entre a entrada e a resposta. Categorias que não são significativamente diferentes (ou seja, que têm um valor p maior que 0,1) são mescladas. Se todas as categorias forem mescladas em uma, as versões originais e derivadas do campo serão excluídas do modelo porque elas não têm nenhum valor como um preditor.

Nível de confiança. Esse é o nível de confiança usado para calcular estimativas de intervalo dos coeficientes do modelo na visualização de Coeficientes. Especifique um valor maior que 0 e menor que 100. O padrão é 95.

Seleção de Modelo

Método de seleção de modelo. Escolha um dos métodos de seleção de modelo (detalhes abaixo) ou **Inclua todos os preditores**, o que simplesmente insere todos os preditores disponíveis como termos modelo dos principais efeitos. Por padrão, **Forward stepwise** é usado.

Seleção de Forward Stepwise. Isso inicia sem nenhum efeito no modelo e inclui e remove efeitos um passo por vez até que nada mais possa ser incluído ou removido de acordo com os critérios stepwise.

- **Critérios para entrada/remoção.** Essa é a estatística usada para determinar se um efeito deve ser incluído ou removido a partir do modelo. **Critério de Informações (AICC)** é baseado na probabilidade do conjunto de treinamento dado o modelo e é ajustado para penalizar modelos excessivamente complexos. **Estatísticas F** são baseadas em um teste estatístico da melhoria no erro de modelo. **R-quadrado ajustado** é baseado no ajuste do conjunto de treinamento e é ajustado para penalizar modelos excessivamente complexos. **Critério de Prevenção ao Super Ajuste (ASE)** é baseado no ajuste (erro quadrático de média ou ASE) do conjunto de prevenção ao super ajuste. O conjunto de prevenção ao super ajuste é uma subamostra aleatória de aproximadamente 30% do conjunto de dados original que não é usado para treinar o modelo.

  Se qualquer critério diferente de **Estatísticas F** for escolhido, então, em cada passo o efeito que corresponde ao maior aumento positivo no critério será incluído no modelo. Qualquer efeito no modelo que corresponde a uma diminuição no critério é removido.

  Se **Estatísticas F** for escolhido como o critério, então, em cada passo o efeito que tiver o menor valor *p* inferior ao limite especificado, **Incluir os efeitos com valores p inferiores a**, será incluído no modelo. O padrão é 0,05. Qualquer efeito no modelo com um valor *p* maior que o limite especificado, **Remover efeitos com valores p maiores que**, será removido. O padrão é 0,10.
- **Customizar número máximo de efeitos no modelo final.** Por padrão, todos os efeitos disponíveis podem ser inseridos no modelo. Como alternativa, se o algoritmo stepwise terminar um passo com o número máximo de efeitos, o algoritmo para com o conjunto atual de efeitos.
- **Customizar número máximo de passos.** O algoritmo stepwise para após um determinado número de passos. Por padrão, esse é 3 vezes o número de efeitos disponíveis. Como alternativa, especifique um número máximo de número inteiro positivo de passos.

Melhor Seleção de Subconjuntos. Isso verifica "todos possíveis" modelos ou pelo menos um maior subconjunto de possíveis modelos do que forward stepwise, para escolher o melhor de acordo com o melhor critério de subconjuntos. **Critério de Informações (AICC)** é baseado na probabilidade do conjunto de treinamento dado o modelo e é ajustado para penalizar modelos excessivamente complexos. **R-quadrado ajustado** é baseado no ajuste do conjunto de treinamento e é ajustado para penalizar modelos excessivamente complexos. **Critério de Prevenção ao Super Ajuste (ASE)** é baseado no ajuste (erro quadrático de média ou ASE) do conjunto de prevenção ao super ajuste. O conjunto de prevenção ao super ajuste é uma subamostra aleatória de aproximadamente 30% do conjunto de dados original que não é usado para treinar o modelo.

O modelo com o maior valor do critério é escolhido como o melhor modelo.

Nota: A seleção dos melhores subconjuntos é computacionalmente mais intensiva do que a seleção forward stepwise. Quando a seleção de melhores subconjuntos é executada em conjunção com boosting, bagging ou conjuntos de dados muito grandes, isso pode levar consideravelmente mais tempo para construir do que um modelo padrão construído usando a seleção forward stepwise.

Combinações

Essas configurações determinam o comportamento da combinação que ocorre quando boosting, bagging ou conjuntos de dados muito grandes são solicitados em Objetivos. Opções que não se aplicam ao objetivo selecionado são ignoradas.

Bagging e Conjuntos de Dados Muito Grandes. Ao pontuar uma combinação, essa é a regra usada para combinar os valores preditos a partir dos modelos base para calcular o valor de escore de combinação.

- **Regra de combinação padrão para variáveis de resposta contínua.** Valores preditos de combinação para variáveis de resposta contínua podem ser combinados usando a média ou a mediana dos valores preditos a partir dos modelos base.

Observe que quando o objetivo é aprimorar a precisão do modelo, as seleções de regra de combinação são ignoradas. Boosting sempre usa um voto de maioria ponderada para escorar variáveis resposta categórica e uma mediana ponderada para escorar variáveis de resposta contínua.

Boosting e Bagging. Especifique o número de modelos base para construir quando o objetivo for aprimorar a precisão ou a estabilidade do modelo; para bagging, esse é o número de amostras bootstrap. Ele deve ser um número inteiro positivo.

Avançado

Replicar resultados. Configurar uma semente aleatória permite que você replique análises. O gerador de números aleatórios é usado para escolher quais registros estão no conjunto de prevenção ao super ajuste. Especifique um número inteiro ou clique em **Gerar**, o que criará um número inteiro pseudo-aleatório entre 1 e 2147483647, inclusive. O padrão é 54752075.

Opções de Modelo

Salve valores preditos no conjunto de dados. O nome de variável padrão é *PredictedValue*.

Modelo de exportação. Isso grava o modelo em um arquivo `.zip` externo. É possível usar esse arquivo de modelo para aplicar as informações de modelo a outros arquivos de dados para propósitos de escoragem. Especifique um nome do arquivo exclusivo, válido. Se a especificação do arquivo se referir a um arquivo existente, então, o arquivo será sobrescrito.

Sumarização do Modelo

A visualização de Sumarização do Modelo é uma captura instantânea, sumarização de visão rápida do modelo e de seu ajuste.

Tabela. A tabela identifica algumas configurações de modelo de alto nível, incluindo:

- O nome da resposta especificado na guia Campos,
- Se a preparação de dado automático foi executada conforme especificado nas configurações Básicas,
- O método de seleção de modelo e o critério de seleção especificados nas configurações de Seleção de Modelo. O valor do critério de seleção para o modelo final também é exibido e é apresentado em formato menor, que é melhor.

Gráfico. O gráfico exibe a precisão do modelo final, que é apresentada em formato maior, que é melhor. O valor é $100 \times$ O R^2 ajustado para o modelo final.

Preparação de Dado Automático

Essa visualização mostra informações sobre quais campos foram excluídos e como campos transformados foram derivados no passo de preparação de dado automático (ADP). Para cada campo que foi transformado ou excluído, a tabela lista o nome do campo, o seu papel na análise e a ação tomada pelo passo de ADP. Campos são ordenados por ordem alfabética ascendente de nomes de campo. As possíveis ações tomadas para cada campo incluem:

- **Duração da derivação: meses** calcula o tempo decorrido em meses a partir dos valores em um campo que contém datas para a data do sistema atual.

- **Duração da derivação: horas** calcula o tempo decorrido em horas a partir dos valores em um campo que contém horas para o tempo do sistema atual.
- **Mudar nível de medição de contínuo para ordinal** reformula campos contínuos com menos de 5 valores exclusivos como campos ordinais.
- **Mudar nível de medição de ordinal para contínuo** reformula campos ordinais com mais de 10 valores exclusivos como campos contínuos.
- **Aparar valores discrepantes** configura valores de preditores contínuos que estão além de um valor de corte (3 desvios padrão a partir da média) para o valor de corte.
- **Substituir valores omissos** substitui valores omissos de campos nominais com os campos de modo, ordinais pelos campos de mediana e contínuos com a média.
- **Mesclar categorias para maximizar associação com resposta** identifica categorias do preditor "semelhantes" com base no relacionamento entre a entrada e a resposta. Categorias que não são significativamente diferentes (ou seja, que têm um valor *p* maior que 0,05) são mescladas.
- **Excluir preditor de constante / após o tratamento de valor discrepante / após a combinação de categorias** remove preditores que têm um valor único, possivelmente após outras ações de ADP terem sido tomadas.

Importância do Preditor

Normalmente, você desejará focar os seus esforços de modelagem nos campos preditores que importam mais e considerar descartar ou ignorar aqueles que menos importam. O gráfico de importância do preditor ajuda você a executar isso, indicando a importância relativa de cada preditor para estimar o modelo. Como os valores são relativos, a soma dos valores para todos os preditores na exibição é 1,0. A importância do preditor não é relacionada com a precisão do modelo. Ela é relacionada apenas com a importância de cada preditor em fazer uma predição, não se a predição é exata ou não.

Predito por Observado

Isso exibe um gráfico de dispersão categorizado dos valores preditos no eixo vertical pelos valores observados no eixo horizontal. Idealmente, os pontos devem estar em uma linha de 45 graus; essa visualização pode lhe dizer se algum registro foi particularmente previsto de forma inválida pelo modelo.

Resíduos

Isso exibe um gráfico de diagnóstico de resíduos do modelo.

Estilos de gráfico. Existem estilos de exibição diferentes, que são acessíveis a partir da lista suspensa de **Estilos**.

- **Histograma.** Esse é um histograma categorizado dos resíduos estudentizados com uma sobreposição da distribuição normal. Modelos lineares assumem que os resíduos têm uma distribuição normal, assim, o histograma deve, idealmente, aproximar bem a linha suave.
- **Gráfico P-P.** Esse é um gráfico de probabilidade-probabilidade categorizado comparando os resíduos estudentizados a uma distribuição normal. Se a inclinação dos pontos criados em gráfico for menos acentuada que a linha normal, os resíduos mostram maior variabilidade que uma distribuição normal; se a inclinação for mais acentuada, os resíduos mostram menos variabilidade que uma distribuição normal. Se os pontos criados em gráfico tiverem uma curva em formato S, então, a distribuição de resíduos será desviada.

Valores Discrepantes

Essa tabela lista registros que exercem uma influência indevida sobre o modelo e exibe o ID de registro (se especificado na guia Campos), o valor de resposta e a distância de Cook. A distância de Cook é uma medida de quanto os resíduos de todos os registros mudariam se um determinado registro fosse excluído do cálculo dos coeficientes do modelo. Uma distância de Cook grande indica que excluir um registro muda os coeficientes substancialmente e deve, portanto, ser considerada influente.

Registros influentes devem ser examinados cuidadosamente para determinar se é possível fornecer a eles menos peso na estimativa do modelo, truncar os valores discrepantes para algum limite aceitável ou remover os registros influentes completamente.

Efeitos

Essa visualização exibe o tamanho de cada efeito no modelo.

Estilos. Existem estilos de exibição diferentes, que são acessíveis a partir da lista suspensa de **Estilos**.

- **Diagrama.** Esse é um gráfico no qual os efeitos são ordenados da parte superior até a inferior diminuindo a importância do preditor. Linhas de conexão no diagrama são ponderadas com base na significância do efeito, com maior largura da linha correspondendo a efeitos mais significativos (valores *p* menores). Passar o mouse sobre uma linha de conexão revela uma dica de ferramenta que mostra o valor *p* e a importância do efeito. Este é o padrão.
- **Tabela.** Essa é uma tabela ANOVA para o modelo global e os efeitos do modelo individuais. Os efeitos individuais são ordenados da parte superior até a inferior diminuindo a importância do preditor. Observe que, por padrão, a tabela é reduzida para mostrar somente os resultados para o modelo global. Para ver os resultados para os efeitos do modelo individuais, clique na célula de **Modelo Corrigido** na tabela.

Importância do preditor. Há uma régua de controle de Importância do Preditor que controla quais preditores são mostrados na visualização. Isso não muda o modelo, mas simplesmente permite que você se concentre nos preditores mais importantes. Por padrão, os 10 efeitos principais são exibidos.

Significância. Há uma régua de controle de Significância que controla adicionalmente quais efeitos são mostrados na visualização, além daqueles mostrados com base na importância do preditor. Efeitos com valores de significância maiores que o valor da régua de controle são ocultados. Isso não muda o modelo, mas simplesmente permite que você se concentre nos efeitos mais importantes. Por padrão o valor é 1,00, de forma que nenhum efeito seja filtrado com base na significância.

Coeficientes

Essa visualização exibe o valor de cada coeficiente no modelo. Observe que fatores (preditores categóricos) são codificados pelo indicador dentro do modelo, de forma que **efeitos** contendo fatores geralmente terão múltiplos **coeficientes** associados; um para cada categoria, exceto a categoria correspondente ao parâmetro (referência) redundante.

Estilos. Existem estilos de exibição diferentes, que são acessíveis a partir da lista suspensa de **Estilos**.

- **Diagrama.** Esse é um gráfico que exibe o intercepto primeiro e, em seguida, ordena efeitos da parte superior até a inferior diminuindo a importância do preditor. Dentro de efeitos que contêm fatores, coeficientes são ordenados por ordem ascendente de valores dos dados. Linhas de conexão no diagrama são coloridas com base no sinal do coeficiente (veja a chave do diagrama) e ponderadas com base na significância do coeficiente, com maior largura da linha correspondendo a coeficientes mais significativos (valores *p* menores). Passar o mouse sobre uma linha de conexão revela uma dica de ferramenta que mostra o valor do coeficiente, o seu valor *p* e a importância do efeito com o qual o parâmetro está associado. Esse é o estilo padrão.

- **Tabela.** Isso mostra os valores, testes de significância e intervalos de confiança para os coeficientes do modelo individuais. Após o intercepto, os efeitos são ordenados da parte superior até a inferior diminuindo a importância do preditor. Dentro de efeitos que contêm fatores, coeficientes são ordenados por ordem ascendente de valores dos dados. Observe que, por padrão, a tabela é reduzida para mostrar somente o coeficiente, a significância e a importância de cada parâmetro de modelo. Para ver o erro padrão, a estatística *t* e o intervalo de confiança, clique na célula de **Coeficiente** na tabela. Passar o mouse sobre o nome de um parâmetro de modelo na tabela revela uma dica de ferramenta que mostra o nome do parâmetro, o efeito com o qual o parâmetro está associado e (para preditores categóricos), os rótulos de valor associados com o parâmetro de modelo. Isso pode ser especialmente útil para ver as novas categorias criadas quando a preparação de dado automático mescla categorias semelhantes de um preditor categórico.

Importância do preditor. Há uma régua de controle de Importância do Preditor que controla quais preditores são mostrados na visualização. Isso não muda o modelo, mas simplesmente permite que você se concentre nos preditores mais importantes. Por padrão, os 10 efeitos principais são exibidos.

Significância. Há uma régua de controle de Significância que controla adicionalmente quais coeficientes são mostrados na visualização, além daqueles mostrados com base na importância do preditor. Coeficientes com valores de significância maiores que o valor da régua de controle são ocultados. Isso não muda o modelo, mas simplesmente permite que você se concentre nos coeficientes mais importantes. Por padrão o valor é 1,00, de forma que nenhum coeficiente seja filtrado com base na significância.

Médias Estimadas

Esses são gráficos exibidos para preditores significativos. O gráfico exibe o valor estimado pelo modelo da resposta no eixo vertical para cada valor do preditor no eixo horizontal, mantendo todos os outros preditores constantes. Ele fornece uma visualização útil dos efeitos dos coeficientes de cada preditor na resposta.

Nota: se nenhum preditor for significativo, nenhuma média estimada será produzida.

Sumarização de Construção de Modelo

Quando um algoritmo de seleção de modelo diferente de **Nenhum** for escolhido nas configurações de Seleção de Modelo, isso fornece alguns detalhes do processo de construção de modelo.

Forward stepwise. Quando forward stepwise for o algoritmo de seleção, a tabela exibirá os últimos 10 passos no algoritmo stepwise. Para cada passo, o valor do critério de seleção e os efeitos no modelo nesse passo são mostrados. Isso lhe dá uma noção de quanto cada passo contribui com o modelo. Cada coluna permite que você ordene as linhas, de forma que seja possível ver mais facilmente quais efeitos estão no modelo em um passo específico.

Melhores subconjuntos. Quando melhores subconjuntos for o algoritmo de seleção, a tabela exibirá os 10 modelos principais. Para cada modelo, o valor do critério de seleção e os efeitos no modelo são mostrados. Isso lhe dá uma noção da estabilidade dos modelos principais; se eles tendem a ter muitos efeitos semelhantes com algumas diferenças, então, é possível estar bastante confiante no modelo "principal"; se eles tendem a ter efeitos muito diferentes, então, alguns dos efeitos podem ser muito semelhantes e devem ser combinados (ou um removido). Cada coluna permite que você ordene as linhas, de forma que seja possível ver mais facilmente quais efeitos estão no modelo em um passo específico.

Capítulo 16. Regressão linear

A Regressão Linear estima os coeficientes da equação linear, envolvendo uma ou mais variáveis independentes, que melhor preveem o valor da variável dependente. Por exemplo, é possível tentar prever o total de vendas anual de um vendedor (a variável dependente) a partir de variáveis independentes como idade, educação e anos de experiência.

Exemplo. O número de jogos vencidos por um time de basquete em uma temporada está relacionado ao número médio de pontos que a equipe marca por jogo? Um gráfico de dispersão indica que essas variáveis estão linearmente relacionadas. O número de jogos vencidos e o número médio de pontos marcados pelo adversário também estão linearmente relacionados. Essas variáveis possuem um relacionamento negativo. Conforme o número de jogos vencidos aumenta, o número médio de pontos marcados pelo adversário diminui. Com a regressão linear, é possível modelar o relacionamento dessas variáveis. Um bom modelo pode ser utilizado para prever quantos jogos os times irão vencer.

Estatísticas. Para cada variável: número de casos válidos, média e desvio padrão. Para cada modelo: coeficientes de regressão, matriz de correlações, correlações de parte ou parciais, R múltiplo, R^2, R^2 ajustado, mudança em R^2, erro padrão da estimativa, tabela de análise de variância, valores preditos e resíduos. Além disso, intervalos de confiança de 95% para cada coeficiente de regressão, matriz de variância-covariância, fator de inflação de variância, tolerância, teste de Durbin-Watson, medidas de distância (de Mahalanobis, Cook e valores de ponto de alavanca), DfBeta, DfFit, intervalos de predição e informações de diagnóstico entre casos. Gráficos: de dispersão, gráficos parciais, histogramas e gráficos de probabilidade normal.

Considerações de Dados da Regressão Linear

Dados. As variáveis dependentes e independentes devem ser quantitativas. Variáveis categóricas, como religião, campo de estudo principal ou região de residência, precisam ser recodificadas para variáveis binárias (simuladas) ou para outros tipos de variáveis de contraste.

Suposições. Para cada valor da variável independente, a distribuição da variável dependente deve ser normal. A variância da distribuição da variável dependente deve ser constante para todos os valores da variável independente. O relacionamento entre a variável dependente e cada variável independente deve ser linear, e todas as observações devem ser independentes.

Para Obter uma análise de Regressão Linear

1. Nos menus, escolha:

   Analisar > **Regressão** > **Linear...**
2. Na caixa de diálogo Regressão Linear, selecione uma variável dependente numérica.
3. Selecione uma ou mais variáveis independentes numéricas.

Como opção, você pode:

- Agrupar variáveis independentes em blocos e especificar métodos de entrada diferentes para diferentes subconjuntos de variáveis.
- Escolher uma variável de seleção para limitar a análise a um subconjunto de casos que possuem um ou mais valores específicos para esta variável.
- Selecionar uma variável de identificação do caso para identificar pontos nos gráficos.
- Selecionar uma variável Ponderação de WLS numérica para uma análise de quadrados mínimos ponderados.

WLS. Permite obter um modelo de quadrados mínimos ponderados. Os pontos de dados são ponderados pelo recíproco de suas variâncias. Isso significa que as observações com variâncias grandes causam menos impacto na análise do que as observações associadas a variâncias pequenas. Se o valor da variável de ponderação for zero, negativo ou omisso, o caso será excluído da análise.

Métodos de Seleção de Variável de Regressão Linear

A seleção de método permite especificar como variáveis independentes são inseridas na análise. Usando métodos diferentes, é possível construir uma variedade de modelos de regressão a partir do mesmo conjunto de variáveis.

- *Inserir (Regressão)*. Um procedimento para seleção de variáveis em que todas as variáveis em um bloco são inseridas em um único passo.
- *Stepwise*. Em cada passo, a variável independente fora da equação que tiver a menor probabilidade de F será inserida, se essa probabilidade for suficientemente pequena. As variáveis que já estiverem na equação de regressão serão removidas se a probabilidade de F for suficientemente grande. O método finalizará quando não houver mais variáveis elegíveis para inclusão ou remoção.
- *Remover*. Um procedimento para seleção de variáveis em que todas as variáveis em um bloco são removidas em um único passo.
- *Eliminação Backward*. Um procedimento de seleção de variáveis em que todas as variáveis são inseridas na equação e, em seguida, removidas sequencialmente. A variável com a menor correlação parcial com a variável dependente é considerada primeiro para remoção. Se ela atender ao critério para eliminação, ela será removida. Após a primeira variável ser removida, a variável restante na equação com a menor correlação parcial é a próxima a ser considerada. O procedimento é interrompido quando não houver variáveis na equação que satisfaçam os critérios de remoção.
- *Seleção Forward*. Um Procedimento de seleção de variáveis stepwise em que as variáveis são inseridas sequencialmente no modelo. A primeira variável considerada para entrada na equação é aquela com a maior correlação positiva ou negativa com a variável dependente. Esta variável será inserida na equação somente se ela satisfizer os critérios para a entrada. Se a primeira variável for inserida, a variável independente que não estiver na equação e que possuir a maior correlação parcial é considerada a próxima. O procedimento é interrompido quando não houver variáveis que atendam ao critério de entrada.

Os valores de significância em sua saída baseiam-se no ajuste de um modelo único. Portanto, os valores de significância são geralmente inválidos quando um método stepwise (stepwise, forward ou backward) é utilizado.

Todas as variáveis devem transmitir o critério de tolerância para serem inseridas na equação, independentemente do método de entrada especificado. O nível de tolerância padrão é 0,0001. Além disso, uma variável não será inserida se ela fizer com que a tolerância de outra variável já no modelo caia abaixo do critério de tolerância.

Todas as variáveis independentes selecionadas são incluídas em um modelo de regressão único. Entretanto, é possível especificar métodos de entrada diferentes para subconjuntos diferentes de variáveis. Por exemplo, é possível inserir um bloco de variáveis no modelo de regressão utilizando a seleção stepwise e um segundo bloco utilizando a seleção forward. Para incluir um segundo bloco de variáveis no modelo de regressão, clique em **Avançar**.

Regra do Conjunto de Regressão Linear

Os casos definidos pela regra de seleção são incluídos na análise. Por exemplo, se selecionar uma variável, escolher **igual a** e digitar 5 para o valor, então apenas os casos para os quais a variável selecionada tiver um valor igual a 5 serão incluídos na análise. Um valor da sequência de caracteres também é permitido.

Gráficos de Regressão Linear

Os gráficos podem ajudar na validação das suposições de normalidade, linearidade e igualdade das variâncias. Os gráficos também são úteis para detectar valores discrepantes, observações incomuns e casos influentes. Após salvá-los como novas variáveis, valores preditos, resíduos e outras informações de diagnóstico estão disponíveis no Editor de Dados para construir gráficos com as variáveis independentes. Os seguintes gráficos estão disponíveis:

Gráficos de dispersão. É possível representar quaisquer dois dos seguintes itens: a variável dependente, valores preditos padronizados, resíduos padronizados, resíduos excluídos, valores preditos ajustados, resíduos Estudentizados ou resíduos excluídos Estudentizados. Represente os resíduos padronizados com relação aos valores preditos padronizados para verificar a linearidade e a igualdade das variâncias.

Lista de variável de origem. Lista a variável dependente (DEPENDNT) e as variáveis preditas e residuais a seguir: Valores preditos padronizados (*ZPRED), Resíduos padronizados (*ZRESID), Resíduos excluídos (*DRESID), Valores preditos ajustados (*ADJPRED), Resíduos estudentizados (*SRESID), Resíduos excluídos estudentizados (*SDRESID).

Produzir todos os gráficos parciais. Exibe gráficos de dispersão de resíduos de cada variável independente e os resíduos da variável dependente quando ambas as variáveis são regredidas separadamente do restante das variáveis independentes. Pelo menos duas variáveis independentes devem estar na equação para que um gráfico parcial seja produzido.

Gráficos de Resíduos Padronizados. É possível obter histogramas de resíduos padronizados e gráficos de probabilidade normal comparando a distribuição de resíduos padronizados com uma distribuição normal.

Se quaisquer gráficos forem solicitados, estatísticas básicas serão exibidas para os valores preditos padronizados e resíduos padronizados (***DRESID** e ***ZRESID**).

Regressão Linear: Salvando novas variáveis

É possível salvar valores preditos, resíduos e outras estatísticas úteis para informações de diagnósticos. Cada seleção inclui uma ou mais novas variáveis no seu arquivo de dados ativos.

Valores Preditos. Valores que o modelo de regressão prevê para cada caso.

- *Não padronizado*. O valor que o modelo prediz para a variável dependente.
- *Padronizado*. Uma transformação de cada valor predito em seu formato padronizado. Ou seja, o valor predito médio é subtraído do valor predito, e a diferença é dividida pelo desvio padrão dos valores preditos. Os valores preditos padronizados possuem uma média de 0 e um desvio padrão de 1.
- *Ajustado*. O valor predito para um caso quando esse caso é excluído do cálculo dos coeficientes de regressão.
- *elemento de suporte de predições de média*. Erros padrão de valores preditos. Uma estimativa do desvio padrão do valor médio da variável dependente para os casos que tiverem os mesmos valores das variáveis independentes.

Distâncias. Medidas para identificar casos com combinações de valores incomuns para as variáveis independentes e casos que podem ter um grande impacto no modelo de regressão.

- *Mahalanobis*. A medida do quanto os valores de um caso nas variáveis independentes diferem da média de todos os casos. Uma distância de Mahalanobis grande identifica um caso como tendo valores extremos em uma ou mais variáveis independentes.
- *de Cook*. Uma medida do quanto os resíduos de todos os casos seriam alterados se um caso específico fosse excluído do cálculo dos coeficientes de regressão. Um D de Cook grande indica que excluir um caso do cálculo das estatísticas de regressão altera os coeficientes substancialmente.

- *REMOVE*. Mede a influência de um ponto no ajuste da regressão. O ponto de alavanca centralizado varia de 0 (nenhuma influência sobre o ajuste) a (N-1)/N.

Intervalos de Predição. Os limites superior e inferior para intervalos de predição de média e individuais.

- *Média*. Limites inferior e superior (duas variáveis) para o intervalo de predição da resposta média predita.
- *Individual*. Limites inferior e superior (duas variáveis) para o intervalo de predição da variável dependente para um caso único.
- *Intervalo de Confiança*. Insira um valor entre 1 e 99,99 para especificar o nível de confiança para os dois Intervalos de Predição. Média ou Individual deve ser selecionado antes de inserir esse valor. Os valores típicos do intervalo de confiança são 90, 95, e 99.

Residuais. O valor real da variável dependente menos o valor predito pela equação de regressão.

- *Não padronizado*. A diferença entre um valor observado e o valor predito pelo modelo
- *Padronizado*. O resíduo dividido por uma estimativa do seu desvio padrão. Resíduos padronizados, também conhecidos como resíduos de Pearson, possuem uma média de 0 e um desvio padrão de 1.
- *Estudentizado*. O resíduo dividido por uma estimativa do seu desvio padrão que varia de caso para caso, dependendo da distância dos valores de cada caso nas variáveis independentes das médias das variáveis independentes.
- *Excluído*. O resíduo para um caso quando esse caso é excluído do cálculo dos coeficientes de regressão. Ele é a diferença entre o valor da variável dependente e o valor predito ajustado.
- *Estudentizado excluído*. O resíduo excluído de um caso dividido pelo seu erro padrão. A diferença entre um resíduo estudentizado excluído e seu resíduo Estudentizado associado indica quanta diferença a eliminação de um caso faz em sua própria predição.

Estatísticas de Influência. A mudança nos coeficientes de regressão (DfBeta[s]) e nos valores preditos (DfFit) que resulta da exclusão de um caso específico. Valores DfBetas e DfFit padronizados também estão disponíveis junto da razão de covariância.

- *DfBeta(s)*. A diferença no valor beta é a mudança no coeficiente de regressão resultante da exclusão de um caso específico. Um valor é calculado para cada termo no modelo, incluindo a constante.
- *DfBeta Padronizado*. A diferença padronizada no valor beta. A mudança no coeficiente de regressão resultante da exclusão de um caso específico. Você pode querer analisar casos com valores absolutos maiores que 2 divididos pela raiz quadrada de N, em que N é o número de casos. Um valor é calculado para cada termo no modelo, incluindo a constante.
- *DfFit*. A diferença no valor de ajuste é a mudança no valor predito que resulta da exclusão de um caso específico.
- *DfFit Padronizado*. Diferença padronizada no valor de ajuste. A mudança no valor predito resultante da exclusão de um caso específico. Você pode querer examinar valores padronizados que, em valor absoluto, excedem 2 vezes a raiz quadrada de p/N, em que p é o número paramétricas no modelo e N é o número de casos.
- *Razão de covariância*. A razão do determinante da matriz de covariâncias com um caso específico excluído do cálculo dos coeficientes de regressão com o determinante da matriz de covariâncias com todos os casos incluídos. Se a razão estiver próxima de 1, o caso não altera significativamente a matriz de covariâncias.

Estatísticas de coeficiente. Salva os coeficientes de regressão em um conjunto de dados ou em um arquivo de dados. Conjuntos de dados estão disponíveis para uso subsequente na mesma sessão, mas não são salvos como arquivos, a menos que sejam salvos explicitamente antes do término da sessão. Os nomes do conjunto de dados devem estar de acordo com as regras de nomenclatura de variáveis.

Exportar informações de modelo em arquivo XML. As estimativas paramétrica e (opcionalmente) suas covariâncias são exportadas no arquivo especificado em formato XML (PMML). É possível usar esse arquivo de modelo para aplicar as informações de modelo a outros arquivos de dados para propósitos de escoragem.

Estatísticas de Regressão Linear

As seguintes estatísticas estão disponíveis:

Coeficientes de Regressão. Estimativas exibe o coeficiente de regressão *B*, o erro padrão de *B*, coeficiente padronizado beta, o valor *t* para *B*, e o nível de significância *t* de dois fatores. **Intervalos de confiança** exibem intervalos de confiança com o nível especificado de confiança para cada coeficiente de regressão ou uma matriz de covariâncias. **Matriz de covariâncias** exibe uma matriz de variâncias-covariâncias de coeficientes de regressão com covariâncias fora da diagonal e variâncias na diagonal. Uma matriz de correlações também é exibida.

Ajuste do modelo. As variáveis inseridas e removidas do modelo são listadas, e as seguintes estatísticas de Qualidade do ajuste são exibidas: *R* múltiplos, R^2 e R^2 ajustado, erro padrão da estimativa e uma tabela de análise de variância.

Alteração de R quadrado. A mudança na estatística R^2 que é produzida ao incluir ou excluir uma variável independente. Se a alteração de R^2 associada a uma variável for grande, isso significa que a variável é uma boa preditora da variável dependente.

Descritivos. Fornece o número de casos válidos, a média e o desvio padrão para cada variável na análise. Uma matriz de correlações com um nível de significância de um fator e o número de casos para cada correlação também são exibidos.

Correlação Parcial. A correlação que permanece entre duas variáveis após remover a correlação que é devido à associação mútua delas com as outras variáveis. A correlação entre a variável dependente e uma variável independente quando os efeitos lineares das outras variáveis independentes no modelo tiverem sido removidos de ambas.

Correlação Semiparcial. A correlação entre a variável dependente e uma variável independente quando os efeitos lineares das outras variáveis independentes no modelo tiverem sido removidos da variável independente. Ela está relacionada à mudança no R-quadrado quando uma variável é incluída em uma equação. Às vezes é chamada de correlação semiparcial.

Diagnósticos de colinearidade. A colinearidade (ou multicolinearidade) é a situação indesejável em que uma variável independente é uma função linear de outras variáveis independentes. Autovalores da matriz de produtos cruzados escalados e não centralizados, índices de condição e proporções de decomposição de variância são exibidos junto de fatores de inflação de variância (VIF) e tolerâncias para variáveis individuais.

Resíduos. Exibe o teste de Durbin-Watson para correlação serial das informações de resíduos e de diagnóstico entre casos para os casos que atenderem ao critério de seleção (valores discrepantes acima de *n* desvios padrão).

Opções de regressão linear

As seguintes opções estão disponíveis:

Critérios de Método Avançandos Estas opções se aplicam quando o método de seleção de variáveis forward, backward e stepwise tiver sido especificado. As variáveis podem ser inseridas ou removidas do modelo, dependendo da significância (probabilidade) do valor *F* ou do próprio valor de *F*.

- *Usar Probabilidade de F*. Uma variável será inserida no modelo se o nível de significância de seu valor F for menor que o valor de Entrada e será removida se o nível de significância for maior que o valor de Remoção. A Entrada deve ser menor que Remoção, e ambos os valores devem ser positivos. Para inserir mais variáveis no modelo, aumente o valor de Entrada. Para remover mais variáveis do modelo, diminua o valor de Remoção.
- *Usar Valor F*. Uma variável será inserida no modelo se seu valor F for maior que o valor de Entrada e será removida se o valor F for menor que o valor de Remoção. A Entrada deve ser maior que Remoção, e ambos os valores devem ser positivos. Para inserir mais variáveis no modelo, diminua o valor de Entrada. Para remover mais variáveis do modelo, aumente o valor de Remoção.

Incluir constante na equação. Por padrão, o modelo de regressão inclui um termo constante. Desmarcar esta opção força a regressão por meio da origem, o que raramente é feito. Alguns resultados de regressão por meio da origem não são comparáveis com os resultados de regressão que incluem uma constante. Por exemplo, R^2 não pode ser interpretado de maneira habitual.

Valores omissos. É possível escolher uma das seguintes opções:

- **Excluir listwise dos casos.** Apenas casos com valores válidos para todas as variáveis são incluídos na análise.
- **Excluir casos entre pares.** Casos com dados completos para o par de variáveis que estão sendo correlacionadas são utilizados para calcular o coeficiente de correlação no qual a análise de regressão é baseada. Graus de liberdade são baseados no *N* mínimo entre pares.
- **Substituir pela média.** Todos os casos são utilizados para cálculos, com a média da variável substituída por observações omissas.

Recursos Adicionais do Comando REGRESSION

O idioma da sintaxe de comando também permite:

- Gravar uma matriz de correlações ou ler uma matriz no lugar dos dados brutos para obter sua análise de regressão (com o subcomando `MATRIX`).
- Especificar níveis de tolerância (com o subcomando `CRITERIA`).
- Obter diversos modelos para as mesmas variáveis dependentes ou diferentes (com os subcomandos `METHOD` e `DEPENDENT`).
- Obter estatísticas adicionais (com os subcomandos `DESCRIPTIVES` e `STATISTICS`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 17. Regressão ordinal

Regressão Ordinal lhe permite modelar a dependência de uma resposta ordinal politômica em um conjunto de preditores, que podem ser fatores ou covariáveis. O design de Regressão Ordinal é baseado na metodologia de McCullagh (1980, 1998) e o procedimento é referido como `PLUM` na sintaxe.

A análise de regressão linear padrão envolve minimizar a soma das diferenças quadráticas entre uma variável de resposta (dependente) e uma combinação ponderada de variáveis preditoras (independentes). Os coeficientes estimados refletem como as mudanças feitas nos preditores afetam a resposta. A resposta é considerada como numérica, no sentido de que mudanças no nível da resposta são equivalentes por todo o intervalo da resposta. Por exemplo, a diferença de altura entre uma pessoa de 1,50 m de altura e uma pessoa de 1,40 m de altura é de 10 cm, que tem o mesmo significado que a diferença de altura entre uma pessoa de 2,10 m de altura e uma pessoa de 2 m de altura. Esses relacionamentos não necessariamente se mantêm para variáveis ordinais, em que a escolha e o número de categorias de resposta podem ser bastante arbitrários.

Exemplo. Regressão Ordinal poderia ser usada para estudar a reação do paciente à dosagem de medicamentos. As possíveis reações podem ser classificadas como *nenhuma*, *suave*, *moderada* ou *grave*. A diferença entre uma reação suave e moderada é difícil ou impossível de quantificar e baseia-se na percepção. Além disso, a diferença entre uma resposta suave e moderada pode ser maior ou menor que a diferença entre uma resposta moderada e grave.

Estatísticas e gráficos. Frequências observadas e esperadas e frequências acumulativas, resíduos de Pearson para frequências e frequências acumulativas, probabilidades observadas e esperadas, probabilidades acumulativas observadas e esperadas de cada categoria de resposta por padrão de covariável, matrizes de correlação e de covariâncias assintóticas de estimativas paramétrica, qui-quadrado de Pearson e qui-quadrado de razão de verossimilhança, estatísticas de qualidade de ajuste, histórico de iteração, teste de suposição de linhas paralelas, estimativas dos parâmetro, erros padrão, intervalos de confiança e estatísticas R^2 de Cox e Snell, de Nagelkerke e McFadden.

Considerações de Dados de Regressão Ordinal

Dados. A variável dependente é considerada como ordinal e pode ser numérica ou sequência de caracteres. A ordenação é determinada pela ordenação dos valores da variável dependente em ordem ascendente. O valor mais baixo define a primeira categoria. Variáveis de fator são consideradas categóricas. Variáveis covariáveis devem ser numéricas. Observe que usar mais de uma covariável contínua pode facilmente resultar na criação de uma tabela de probabilidades de célula muito grande.

Suposições. Somente uma variável de resposta é permitida e ela deve ser especificada. Além disso, para cada padrão distinto de valores ao longo das variáveis independentes, as respostas são consideradas variáveis multinomiais independentes.

Procedimentos relacionados. A regressão logística nominal usa modelos semelhantes para variáveis dependentes nominais.

Obtendo uma Regressão Ordinal

1. Nos menus, escolha:
   Analisar > **Regressão** > **Ordinal...**
2. Selecione uma variável dependente.
3. Clique em **OK**.

Opções de Regressão Ordinal

A caixa de diálogo Opções lhe permite ajustar parâmetros usados no algoritmo de estimação iterativa, escolher um nível de confiança para as suas estimativas paramétrica e selecionar uma função de ligação.

Iterações. É possível customizar o algoritmo iterativo.

- **Máximo de iterações.** Especifique um número inteiro não negativo. Se 0 for especificado, o procedimento retorna as estimativas iniciais.
- **Máximo de redução para metade do passo.** Especifique um número inteiro positivo.
- **Convergência de log da verossimilhança.** O algoritmo parará se a mudança absoluta ou relativa no log da verossimilhança for menor que esse valor. O critério não será usado se 0 for especificado.
- **Convergência de Parâmetro.** O algoritmo parará se a mudança absoluta ou relativa em cada uma das estimativas paramétrica for menor que esse valor. O critério não será usado se 0 for especificado.

Intervalo de confiança. Especifique um valor maior ou igual a 0 e menor que 100.

Delta. O valor incluído em frequências de célula zero. Especifique um valor não negativo menor que 1.

Tolerância à singularidade. Usado para verificar se há preditores altamente dependentes. Selecione um valor a partir da lista de opções.

Função de ligação. A função de ligação é uma transformação da probabilidades acumulativas que permite estimação do modelo. As cinco funções de link a seguir estão disponíveis.

- **Logit.** $f(x)=\log(x\ /\ (1-x))$. Geralmente usada para categorias uniformemente distribuídas.
- **Log-log complementar.** $f(x)=\log(-\log(1-x))$. Geralmente usada quando categorias superiores são mais prováveis.
- **Log-log negativo.** $f(x)=-\log(-\log(x))$. Geralmente usada quando categorias inferiores são mais prováveis.
- **Probito.** $f(x)=\Phi^{-1}(x)$. Geralmente usada quando a variável latente é normalmente distribuída.
- **Cauchit (Cauchy inverso).** $f(x) = \tan(\pi\ (x - 0.5))$. Geralmente usada quando a variável latente possui muitos valores extremos.

Saída de Regressão Ordinal

A caixa de diálogo Saída permite produzir tabelas para exibição no Visualizador e salvar variáveis no arquivo de trabalho.

Exibição. Produz tabelas para:

- **Imprimir histórico de iteração.** As estimativas de log da verossimilhança e paramétrica são impressas para a frequência de iteração de impressão especificada. A primeira e última iterações são sempre impressas.
- **Estatísticas de qualidade de ajuste.** As estatísticas de Pearson e de qui-quadrado de razão de verossimilhança. Elas são calculadas com base na classificação especificada na lista de variáveis.
- **Estatística de sumarização.** Estatísticas R^2 de Cox e Snell, de Nagelkerke e de McFadden.
- **Estimativas paramétrica.** Estimativas paramétrica, erros padrão e intervalos de confiança.
- **Correlação assintótica de estimativas paramétrica.** Matriz de correlações de estimativas paramétrica.
- **Covariância assintótica de estimativas paramétrica.** Matriz de covariâncias de estimativas paramétrica.
- **Informações de célula.** Frequências observadas e esperadas e frequências acumulativas, resíduos de Pearson para frequências e frequências acumulativas, probabilidades observadas e esperadas, probabilidades acumulativas observadas e esperadas de cada categoria de resposta por padrão de covariável. Observe que para modelos com muitos padrões de covariável (por exemplo, modelos com covariáveis contínuas), essa opção pode gerar uma tabela muito grande e pesada.

- **Teste de linhas paralelas.** Teste da hipótese de que parâmetros de localização são equivalentes ao longo dos níveis da variável dependente. Isso está disponível somente para o modelo somente para local.

Variáveis salvas. Salva as variáveis a seguir no arquivo de trabalho:

- **Probabilidades de resposta estimada.** Probabilidades estimadas de modelo de classificação de um padrão de covariável/fator nas categorias de resposta. Há tantas probabilidades quanto o número de categorias de resposta.
- **Categoria predita.** A categoria de resposta que tem a probabilidade máxima estimada para um padrão de covariável/fator.
- **Probabilidade de categoria predita.** Probabilidade estimada de classificação de um padrão de covariável/fator na categoria predita. Essa probabilidade também é o máximo das probabilidades estimadas do padrão de covariável/fator.
- **Probabilidade de categoria real.** Probabilidade estimada de classificação de um padrão de covariável/fator na categoria real.

Imprimir log da verossimilhança. Controla a exibição do log da verossimilhança. **Incluir a constante de multinômio** lhe dá o valor integral da probabilidade. Para comparar os seus resultados ao longo de produtos que não incluem a constante, é possível escolher excluí-la.

Modelo de Localização de Regressão Ordinal

A caixa de diálogo Localização permite especificar o modelo de localização para a sua análise.

Especificar modelo. Um modelo dos principais efeitos contém os efeitos principais de covariável e fator, mas nenhum efeito de interação. É possível criar um modelo customizado para especificar subconjuntos de interações de fatores ou interações de covariáveis.

Fatores/covariáveis. Os fatores e covariáveis são listados.

Modelo de localização. O modelo depende dos efeitos principais e dos efeitos de interação que você selecionar.

Termos de compilação

Para os fatores e covariáveis selecionados:

Interação. Cria o termo de interação de nível mais alto de todas as variáveis selecionadas. Esse é o padrão.

Efeitos principais. Cria um termo dos principais efeitos para cada variável selecionada.

Todos os 2 fatores. Cria todas as interações de dois fatores possíveis das variáveis selecionadas.

Todas 3 fatores. Cria todas as interações de três fatores das variáveis selecionadas.

Todos os 4 fatores. Cria todas as interações de quatro fatores possíveis das variáveis selecionadas.

Todos os 5 fatores. Cria todas as interações de cinco fatores possíveis das variáveis selecionadas.

Modelo de Escala de Regressão Ordinal

A caixa de diálogo Escala permite especificar o modelo de escala para a sua análise.

Fatores/covariáveis. Os fatores e covariáveis são listados.

Modelo de escala. O modelo depende dos efeitos principais e de interação que você selecionar.

Termos de compilação

Para os fatores e covariáveis selecionados:

Interação. Cria o termo de interação de nível mais alto de todas as variáveis selecionadas. Esse é o padrão.

Efeitos principais. Cria um termo dos principais efeitos para cada variável selecionada.

Todos os 2 fatores. Cria todas as interações de dois fatores possíveis das variáveis selecionadas.

Todas 3 fatores. Cria todas as interações de três fatores das variáveis selecionadas.

Todos os 4 fatores. Cria todas as interações de quatro fatores possíveis das variáveis selecionadas.

Todos os 5 fatores. Cria todas as interações de cinco fatores possíveis das variáveis selecionadas.

Recursos Adicionais do Comando PLUM

É possível customizar sua Regressão Ordinal se colar suas seleções em uma janela de sintaxe e editar a sintaxe de comando `PLUM` resultante. O idioma da sintaxe de comando também permite:

- Criar testes de hipótese customizados ao especificar hipóteses nulas como combinações lineares paramétricas.

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 18. Curva de Estimação

O procedimento Curva de Estimação produz estatísticas de regressão de curva de estimação e gráficos relacionados para 11 modelos de regressão de curva de estimação diferentes. Um modelo separado é produzido para cada variável dependente. Também é possível salvar valores preditos, resíduos e intervalos de predição como novas variáveis.

Exemplo. Um provedor de serviços da Internet monitora a porcentagem de tráfego de e-mail infectado por vírus em suas redes ao longo do tempo. Um gráfico de dispersão revela que o relacionamento é não linear. É possível ajustar um modelo quadrático ou cúbico nos dados e verificar a validade das suposições e a Qualidade do ajuste do modelo.

Estatísticas. Para cada modelo: coeficientes de regressão, *R* múltiplo, R^2, R^2 ajustado, erro padrão da estimativa, tabela de análise de variância, valores preditos, resíduos e intervalos de predição. Modelos: linear, logarítmico, inverso, quadrático, cúbico, potência, composto, curva S, logístico, crescimento e exponencial.

Considerações de Dados de Curva de Estimação

Dados. As variáveis dependentes e independentes devem ser quantitativas. Se selecionar **Tempo** a partir do conjunto de dados ativo como a variável independente (ao invés de selecionar uma variável), o procedimento Curva de Estimação gera uma variável de tempo em que o período de tempo entre os casos é uniforme. Se **Tempo** for selecionado, a variável dependente deverá ser uma medida de séries temporais. A análise de séries temporais requer uma estrutura de arquivo de dados na qual cada caso (linha) representa um conjunto de observações em um tempo diferente, e o período de tempo entre os casos é uniforme.

Suposições. Verifique seus dados graficamente para determinar como as variáveis independentes e dependentes estão relacionadas (linearmente, exponencialmente, etc.). Os resíduos de um bom modelo devem ser distribuídos aleatoriamente e normais. Se um modelo linear for utilizado, as suposições a seguir devem ser atendidas: Para cada valor da variável independente, a distribuição da variável dependente deve ser normal. A variância da distribuição da variável dependente deve ser constante para todos os valores da variável independente. O relacionamento entre a variável dependente e a variável independente deve ser linear, e todas as observações devem ser independentes.

Para Obter uma Curva de Estimação

1. Nos menus, escolha:

   Analisar > **Regressão** > **Curva de Estimação...**
2. Selecione uma ou mais variáveis dependentes. Um modelo separado é produzido para cada variável dependente.
3. Selecione uma variável independente (ou selecione uma variável no conjunto de dados ativo ou selecione **Tempo**).
4. Opcionalmente:

- Selecione uma variável para rotular casos em gráficos de dispersão. Para cada ponto no gráfico de dispersão, é possível utilizar a ferramenta Seleção de Ponto para exibir o valor da variável Rótulo Case.
- Clique em **Salvar** para salvar valores preditos, resíduos e intervalos de predição como novas variáveis.

As seguintes opções também estão disponíveis:

- **Incluir constante na equação.** Estima um termo constante na equação de regressão. A constante é incluída por padrão.

- **Modelos de gráfico.** Representa os valores da variável dependente e cada modelo selecionado com relação à variável independente. Um gráfico separado é produzido para cada variável dependente.
- **Exibir tabela ANOVA.** Exibe uma tabela de análise de variância resumida para cada modelo selecionado.

Modelos de Curva de Estimação

É possível escolher um ou mais modelos de regressão de estimativa de curva. Para determinar qual modelo utilizar, crie um gráfico de seus dados. Se as variáveis aparentarem estar relacionadas linearmente, use um modelo de regressão linear simples. Quando suas variáveis não estiverem linearmente relacionadas, tente transformar seus dados. Quando uma transformação não ajuda, poderá ser necessário um modelo mais complicado. Visualize um gráfico de dispersão de seus dados; se o gráfico for semelhante a uma função matemática conhecida, ajuste os dados para esse tipo de modelo. Por exemplo, se seus dados forem semelhantes a uma função exponencial, utilize um modelo exponencial.

Linear. Modelo cuja equação é Y = b0 + (b1 * t). Os valores de série são modelados como uma função linear de tempo.

Logarítmico. Modelo cuja equação é Y = b0 + (b1 * ln(t)).

Inverso. Modelo cuja equação é Y = b0 + (b1 / t).

Quadrático. Modelo cuja equação Y = b0 + (b1 * t) + (b2 * t**2). O modelo quadrático pode ser usado para modelar uma série que "decola" ou uma série que amortece.

Cúbico. Modelo que é definido pela equação Y = b0 + (b1 * t) + (b2 * t**2) + (b3 * t**3).

Potência. Modelo cuja equação é Y = b0 * (t**b1) ou ln(Y) = ln(b0) + (b1 * ln(t)).

Composto. Modelo cuja equação é Y = b0 * (b1**t) ou ln(Y) = ln(b0) + (ln(b1) * t).

Curva S. Modelo cuja equação é Y = e**(b0 + (b1/t)) ou ln(Y) = b0 + (b1/t).

Logística. Modelo cuja equação é Y = 1 / (1/u + (b0 * (b1**t))) ou ln(1/y-1/u) = ln (b0) + (ln(b1) * t), em que u é o valor limite superior. Após selecionar Logística, especifique o valor limite superior a ser usado na equação de regressão. O valor deve ser um número positivo maior que o maior valor de variável dependente.

Crescimento. Modelo cuja equação é Y = e**(b0 + (b1 * t)) ou ln(Y) = b0 + (b1 * t).

Exponencial. Modelo cuja equação é Y = b0 * (e**(b1 * t)) ou ln(Y) = ln(b0) + (b1 * t).

Salvar na Curva de Estimação

Salvar Variáveis. Para cada modelo selecionado, é possível salvar valores preditos, resíduos (valor observado da variável dependente menos o valor predito do modelo) e intervalos de predição (limites superior e inferior). Os novos nomes de variáveis e rótulos descritivos são exibidos em uma tabela na janela de saída.

Prever Casos. No conjunto de dados ativo, se selecionar **Tempo** ao invés de uma variável como a variável independente, será possível especificar um período de previsão além do término da série temporal. É possível escolher uma das seguintes alternativas:

- **Prever a partir do período de estimação até o último caso.** Prevê valores para todos os casos no arquivo, com base nos casos no período de estimação. O período de estimação, exibido na parte

inferior da caixa de diálogo, é definido com a caixa de subdiálogo Intervalo da opção Selecionar Casos no menu de Dados. Se nenhum período de estimação tiver sido definido, todos os casos serão usados para prever valores.

- **Prever através de.** Prevê valores por meio da data, hora ou número de observação especificado, com base nos casos no período de estimação. Esse recurso pode ser utilizado para prever valores além do último caso na série temporal. As variáveis de data definidas atualmente determinam quais caixas de texto estão disponíveis para especificar o término do período de predição. Se não houver nenhuma variável de data definida, será possível especificar o número de observação final (caso).

Utilize a opção Definir Datas no menu Dados para criar variáveis de data.

Capítulo 19. Regressão por quadrados mínimos parciais

O procedimento Regressão por Quadrados Mínimos Parciais estima modelos de regressão de quadrados mínimos parciais (PLS, também conhecido como "projeção para estrutura latente"). O PLS é uma técnica preditiva que é uma alternativa para a regressão por quadrados mínimos (OLS) ordinária, correlação canônica ou modelagem de equação estrutural, e é particularmente útil quando variáveis preditoras estão altamente correlacionadas ou quando o número de preditores excede o número de casos.

O PLS combina os recursos de análise de componentes principais e de diversas regressões. Ele primeiro extrai um conjunto de fatores latentes que explicam o máximo de covariância possível entre as variáveis independentes e dependentes. Em seguida, um passo de regressão prevê valores das variáveis dependentes utilizando a decomposição das variáveis independentes.

Tabelas. Proporção da variância explicada (por fator latente), ponderações de fatores latentes, carregamentos de fatores latentes, a importância de variável independente na projeção (VIP) e as estimativas paramétrica de regressão (por variável dependente) são todos produzidos por padrão.

Gráficos. Importância de variável na projeção (VIP), escores de fatores, ponderações de fatores para os três primeiros fatores latentes e a distância para o modelo são todos produzidos a partir da guia Opções.

Considerações de Dados de Regressão por Quadrados Mínimos Parciais

Nível de medição. Varáveis dependentes e independentes (preditoras) podem ser de escala, nominais ou ordinais. O procedimento supõe que o nível de medição apropriado foi designado para todas as variáveis, embora seja possível alterar temporariamente o nível de medição para uma variável clicando com o botão direito na variável na lista de variáveis de origem e selecionando um nível de medição no menu pop-up. Variáveis categóricas (nominais ou ordinais) são tratadas de maneira equivalente pelo procedimento.

Codificação de variável categórica. O procedimento recodifica temporariamente variáveis dependentes categóricas utilizando uma codificação um-de-*c* durante o procedimento. Se houver *c* categorias de uma variável, a variável será armazenada como *c* vetores, com a primeira categoria indicada (1,0,...,0), com a próxima categoria (0,1,0,...,0),..., e com a categoria final (0,0,...,0,1). Variáveis dependentes categóricas são representadas utilizando codificação simulada, isto é, simplesmente omite o indicador correspondente para a categoria de referência.

Ponderações de frequência. Valores de ponderação são arredondados para o número inteiro mais próximo antes do uso. Casos com ponderações omissas ou ponderações menores que 0,5 não são utilizados na análise.

Valores omissos. Os valores omissos do usuário e do sistema são tratados como inválidos.

Escalando novamente. Todas as variáveis de modelo são centralizadas e padronizadas, incluindo variáveis indicadoras representando as variáveis categóricas.

Para Obter Regressão por Quadrados Mínimos Parciais

Nos menus, escolha:

Analisar > **Regressão** > **Quadrados Mínimos Parciais ...**

1. Selecione pelo menos uma variável dependente.
2. Selecione pelo menos uma variável independente.

Como opção, você pode:

- Especificar uma categoria de referência para variáveis dependentes categóricas (nominais ou ordinais).
- Especificar uma variável a ser utilizada como um identificador exclusivo para a saída de casos e conjuntos de dados salvos.
- Especificar um limite superior no número de fatores latentes a serem extraídos.

Pré-requisitos

O procedimento Regressão por Quadrados Mínimos Parciais é um comando de extensão do Python que requer o IBM SPSS Statistics - Essentials for Python, que é instalado por padrão com o produto IBM SPSS Statistics. Ele requer também as bibliotecas Python NumPy e SciPy, que estão gratuitamente disponíveis.

Nota: Para os usuários que trabalham no modo de análise distribuída (requer o IBM SPSS Statistics Server), o NumPy e o SciPy devem estar instalados no servidor. Entre em contato com seu administrador de sistema para obter assistência.

Usuários do Windows e Mac

Para Windows e Mac, o NumPy e SciPy devem ser instalados para uma versão separada do Python 2.7 da versão que está instalada com o IBM SPSS Statistics. Se você não tiver uma versão separada do Python 2.7, será possível fazer download dele a partir do http://www.python.org. Em seguida, instale o NumPy e o SciPy para o Python versão 2.7. Os instaladores estão disponíveis a partir do http://www.scipy.org/Download.

Para permitir o uso do NumPy e SciPy, deve-se configurar sua localização do Python para a versão do Python 2.7 na qual você instalou o NumPy e o SciPy. A localização do Python é configurada a partir da guia Localizações de Arquivo no diálogo Opções (Editar > Opções).

Usuários do Linux

Sugerimos que você mesmo faça download da origem e construa o NumPy e o SciPy. A origem está disponível a partir do http://www.scipy.org/Download. É possível instalar o NumPy e o SciPy para a versão Python 2.7 que é instalada com o IBM SPSS Statistics. Ele está no diretório `Python` na localização em que o IBM SPSS Statistics está instalado.

Se optar por instalar o NumPy e o SciPy em uma versão do Python 2.7 diferente da versão que está instalada com o IBM SPSS Statistics, então deve-se configurar sua localização do Python para apontar para essa versão. A localização do Python é configurada a partir da guia Localizações de Arquivo no diálogo Opções (Editar > Opções).

Usuários do Windows e Unix

O NumPy e o SciPy devem ser instalados, no servidor, para uma versão separada do Python 2.7 da versão que está instalada com o IBM SPSS Statistics. Se não houver uma versão separada do Python 2.7 no servidor, então ela poderá ser transferida por download a partir do http://www.python.org. O NumPy e o SciPy para Python 2.7 estão disponíveis a partir do http://www.scipy.org/Download. Para possibilitar o uso do NumPy e SciPy, a localização do Python para o servidor deve ser configurada para a versão do Python 2.7 em que o NumPy e SciPy estão instalados. A localização do Python é configurada a partir do IBM SPSS Statistics Administration Console.

Modelo

Especificar Efeitos do Modelo. Um modelo de efeitos principal contém todos os efeitos principais de fator e covariáveis. Selecione **Customizado** para especificar interações. Deve-se indicar todos os termos a serem incluídos no modelo.

Fatores e Covariáveis. Os fatores e covariáveis são listados.

Modelo. O modelo depende da natureza de seus dados. Após selecionar **Customizado**, é possível selecionar os efeitos e interações principais que forem de interesse em sua análise.

Criar termos

Para os fatores e covariáveis selecionados:

Interação. Cria o termo de interação de nível mais alto de todas as variáveis selecionadas. Esse é o padrão.

Efeitos principais. Cria um termo dos principais efeitos para cada variável selecionada.

Todas 2 fatores. Cria todas as interações de dois fatores possíveis das variáveis selecionadas.

Todas 3 fatores. Cria todas as interações de três fatores possíveis das variáveis selecionadas.

Todas 4 fatores. Cria todas as interações de quatro fatores possíveis das variáveis selecionadas.

Todas 5 fatores. Cria todas as interações de cinco fatores possíveis das variáveis selecionadas.

Opções

A guia Opções permite ao usuário salvar e criar um gráfico de estimativas de modelo para casos individuais, fatores latentes e preditores.

Para cada tipo de dados, especifique o nome de um conjunto de dados. Os nomes dos conjuntos de dados devem ser exclusivos. Se especificar o nome de um conjunto de dados existente, seu conteúdo será substituído; caso contrário, um novo conjunto de dados será criado.

- **Salvar estimativas para casos individuais.** Salva as seguintes estimativas de modelo de casos: valores preditos, residuais, distância para modelo de fator latente e escores dos fatores latentes. Também cria um gráfico dos escores dos fatores latentes.
- **Salvar estimativas para fatores latentes.** Salva carregamentos e ponderações de fatores latentes. Também cria um gráfico das ponderações de fatores latentes.
- **Salvar estimativas para variáveis independentes.** Salva estimativas do parâmetro de regressão e a importância de variável para projeção (VIP). Também cria um gráfico de uma VIP por fator latente.

Capítulo 20. Análise do vizinho mais próximo

A Análise do Vizinho Mais Próximo é um método de classificação de casos com base na sua similaridade com outros casos. Em aprendizado por máquina, ela foi desenvolvida como uma maneira de reconhecer padrões de dados sem requerer uma correspondência exata com nenhum dos padrões ou casos armazenados. Os casos semelhantes estão próximos uns dos outros e os casos dissimilares estão distantes. Portanto, a distância entre dois casos é uma medida de sua dissimilaridade.

Os casos que estiverem próximos uns dos outros são chamados de "vizinhos". Quando um novo caso (validação) é apresentado, sua distância de cada um dos casos no modelo é calculada. As classificações dos casos mais similares – os vizinhos mais próximos – são verificadas e o novo caso é colocado na categoria que contiver o maior número de vizinhos mais próximos.

É possível especificar o número de vizinhos mais próximos a serem examinados; este valor é chamado de *k*.

A análise do vizinho mais próximo também pode ser utilizada para calcular valores para uma variável resposta contínua. Nesta situação, a média ou mediana do valor dos vizinhos mais próximos é utilizada para obter o valor predito para o novo caso.

Considerações de Dados de Análise do Vizinho Mais Próximo

Resposta e recursos. A resposta e os recursos podem ser:

- *Nominal*. Uma variável pode ser tratada como nominal quando seus valores representarem categorias sem ranqueamento intrínseco (por exemplo, o departamento da empresa na qual um funcionário trabalha). Exemplos de variáveis nominais incluem região, código de endereçamento postal e filiação religiosa.
- *Ordinal*. Uma variável pode ser tratada como ordinal quando seus valores representarem categorias com algum ranqueamento intrínseco (por exemplo, níveis de satisfação de serviço de muito insatisfeito para muito satisfeito). Exemplos de variáveis ordinais incluem escores de atitude que representam o grau de satisfação ou de confiança e os escores de classificação de preferência.
- *Escala*. Uma variável pode ser tratada como escala (contínua) quando os seus valores representarem categorias ordenadas com uma métrica significativa, de forma que as comparações de distância entre os valores sejam apropriadas. Exemplos de variáveis de escala incluem idade em anos e rendimento em milhares de dólares.

  As variáveis Nominais e Ordinais são tratadas de forma equivalente pela Análise do Vizinho Mais Próximo. O procedimento supõe que o nível de medição apropriado foi designado para cada variável, no entanto, é possível alterar temporariamente o nível de medição para uma variável clicando com o botão direito na variável na lista de variáveis de origem e selecionando um nível de medição no menu pop-up.

Um ícone ao lado de cada variável na lista de variáveis identifica o nível de medição e o tipo de dados:

Tabela 1. Ícones do nível de medição

	Numérico	**Sequência de caracteres**	**Data**	**Horário**
Escala (Contínua)		n/a		
Ordinal				

Tabela 1. Ícones do nível de medição (continuação)

	Numérico	**Sequência de caracteres**	**Data**	**Horário**
Nominal				

Codificação de variável categórica. O procedimento recodifica temporariamente preditores categóricos e variáveis dependentes utilizando a codificação um-de-*c* durante o procedimento. Se houver *c* categorias de uma variável, então a variável será armazenada como vetores de c, com a primeira categoria denotada (1,0,...,0), com a próxima categoria (0,1,0,...,0), ..., e com a categoria final (0,0,...,0,1).

Esse esquema de codificação aumenta a dimensionalidade do espaço do recurso. Em particular, o número total de dimensões é o número de preditores de escala mais o número de categorias em todos os preditores categóricos. Como resultado, esse esquema de codificação pode levar a um treinamento mais lento. Se o treinamento de seus vizinhos mais próximos continuar muito lentamente, você pode tentar reduzir o número de categorias em seu preditores categóricos ao combinar categorias semelhantes ou descartar casos que possuam categorias extremamente raras antes de executar o procedimento.

Toda a codificação um-de-*c* baseia-se nos dados de treinamento, mesmo se uma amostra de validação estiver definida (consulte "Partições" na página 90). Assim, se a amostra de validação contiver casos com categorias do preditor que não estiverem presentes nos dados de treinamento, então esses casos não serão escorados. Se a amostra de validação contiver casos com categorias de variável dependente que não estiverem presentes nos dados de treinamento, então esses casos serão escorados.

Escalando novamente. Os recursos de escala são normalizados por padrão. Todo o reajuste de escala é executado com base nos dados de treinamento, mesmo se uma amostra de validação estiver definida (consulte "Partições" na página 90). Se especificar uma variável para definir partições, é importante que os recursos tenham distribuições semelhantes entre as amostras de treinamento e de validação. Utilize, por exemplo, o procedimento Explorar para examinar as distribuições entre as partições.

Ponderações de frequência. As ponderações de frequência são ignoradas por este procedimento.

Replicando resultados. O procedimento usa a geração de número aleatório durante a designação aleatória de partições e dobras de validação cruzada. Se desejar replicar seus resultados de modo exato, além de utilizar as mesmas configurações do procedimento, configure um valor semente para o Mersenne Twister (consulte "Partições" na página 90), ou utilize variáveis para definir partições e dobras de validação cruzada.

Para obter uma análise do vizinho mais próximo

Nos menus, escolha:

Analisar > **Classificar** > **Vizinho Mais Próximo...**

1. Especifique um ou mais recursos, que podem ser considerados variáveis ou preditores independentes se houver uma resposta.

   Destino (opcional). Se nenhum destino (variável ou resposta) for especificado, então o procedimento localizará somente os *k* vizinhos mais próximos, e nenhuma classificação ou predição será feita.

   Normalizar recursos de escala. As variáveis normalizadas possuem o mesmo intervalo de valores, que pode melhorar o desempenho do algoritmo de estimação. A normalização ajustada, [2*(x−min)/(max−min)]−1, é usada. Os valores normalizados ajustados estão entre −1 e 1.

   Identificador de caso focal (opcional). Isso permite marcar os casos de interesse especial. Por exemplo, um pesquisador deseja determinar se os escores de teste a partir de um distrito escolar - o caso focal - são comparáveis com aqueles de distritos escolares semelhantes. Ele utiliza a análise do

vizinho mais próximo para localizar os distritos escolares que forem mais semelhantes com relação a um determinado conjunto de recursos. Em seguida, ele compara os escores de teste do distrito escolar focal com aqueles dos vizinhos mais próximos.

Casos focais também podem ser utilizados em estudos clínicos para selecionar casos de controle que forem semelhantes aos casos clínicos. Os casos focais são exibidos nos *k* vizinhos mais próximos e na tabela de distância, nos gráficos de espaço de variáveis, no gráfico de peers e no mapa de quadrante. Informações sobre casos focais são salvas nos arquivos especificados na guia Saída.

Os casos com um valor positivo na variável especificada são tratados como casos focais. É inválido especificar uma variável sem valores positivos.

Rótulo case (opcional). Os casos são rotulados utilizando esses valores no gráfico de espaço da variável, no gráfico de peers e no mapa de quadrante.

Campos com nível de medição desconhecido

O alerta de Nível de Medição é exibido quando o nível de medição para uma ou mais variáveis (campos) no conjunto de dados é desconhecido. Como o nível de medição afeta o cálculo de resultados para este procedimento, todas as variáveis devem ter um nível de medição definido.

Dados de varredura. Lê os dados no conjunto de dados ativo e designa o nível de medição padrão para quaisquer campos com um nível de medição desconhecido atualmente. Se o conjunto de dados for grande, isso poderá demorar algum tempo.

Designar Manualmente. Abre um diálogo que lista todos os campos com um nível de medição desconhecido. É possível utilizar este diálogo para designar o nível de medição para esses campos. Também é possível designar o nível de medição na Visualização de Variável do Editor de Dados.

Como o nível de medição é importante para este procedimento, não é possível acessar o diálogo para executar este procedimento até que todos os campos possuam um nível de medição definido.

Vizinhos

Número de Vizinhos Mais Próximos (k). Especifique o número de vizinhos mais próximos. Observe que utilizar um número maior de vizinhos não resultará necessariamente em um modelo mais preciso.

Se uma resposta for especificada na guia Variáveis, será possível especificar, como alternativa, um intervalo de valores e permitir que o procedimento escolha o "melhor" número de vizinhos dentro desse intervalo. O método de determinação do número de vizinhos mais próximos depende se a seleção de variável é solicitada na guia Recursos.

- Se a seleção de variável estiver em vigor, então a seleção de variável será executada para cada valor de *k* no intervalo solicitado, e o *k*, junto do conjunto de recursos acompanhante, com a menor taxa de erro (ou a menor soma dos quadrados dos erros se a resposta for escalar) será selecionado.
- Se a seleção de variável não estiver em vigor, então a validação cruzada da dobra *V* será utilizada para selecionar o "melhor" número de vizinhos. Consulte a guia Partição para controlar sobre designação de dobras.

Cálculo de Distância. Esta é a métrica utilizada para especificar a métrica de distância utilizada para medir a similaridade de casos.

- **Métrica euclidiana.** A distância entre dois casos, x e y, é a raiz quadrada da soma, em todas as dimensões, das diferenças quadráticas entre os valores para os casos.
- **Métrica de City Block.** A distância entre dois casos é a soma, em todas as dimensões, das diferenças absolutas entre os valores para os casos. Essa métrica também é chamada de distância de Manhattan.

Opcionalmente, se uma resposta for especificada na guia Variáveis, será possível optar por ponderar recursos pela sua importância normalizada ao calcular distâncias. A importância de recurso para um preditor é calculada pela razão entre a taxa de erros ou a soma dos quadrados dos erros do modelo com o preditor removido do modelo com a taxa de erros ou a soma dos quadrados dos erros do modelo completo. A importância normalizada é calculada ao reponderar os valores de importância de variável para que a soma deles seja 1.

Predições para Variável Resposta Escalar. Se uma variável resposta escalar for especificada na guia Variáveis, isso especifica se o valor predito é calculado com base na média ou no valor mediano nos vizinhos mais próximos.

Recursos

A guia Recursos permite solicitar e especificar opções para seleção de variáveis quando uma resposta é especificada na guia Variáveis. Por padrão, todas as variáveis são consideradas para seleção de variável, no entanto, é possível, opcionalmente, selecionar um subconjunto de variáveis para forçar no modelo.

Critério de Parada. Em cada passo, o recurso cuja adição ao modelo resulta no menor erro (calculado como a taxa de erros para uma variável resposta categórica e a soma dos erros quadráticos para uma resposta de escala) é considerado para inclusão no conjunto de modelo. A seleção Forward continua até que a condição especificada seja atendida.

- **Número de recursos especificados.** O algoritmo inclui um número fixo de variáveis além daquelas forçadas no modelo. Especifique um número inteiro positivo. Diminuir valores do número para seleção cria um modelo mais simples, sob risco de perder variáveis importantes. Aumentar valores do número para seleção capturará todas as variáveis importantes, sob risco de incluir eventualmente variáveis que na realidade aumentam o erro do modelo.
- **Mudança mínima na razão de erro absoluto.** O algoritmo para quando a mudança na razão de erro absoluto indicar que o modelo não poderá ser melhorado ainda mais ao incluir mais variáveis. Especifique um número positivo. Diminuir valores da mudança mínima tende a incluir mais recursos, com o risco de incluir recursos que não agregam muito valor no modelo. Aumentar o valor da mudança mínima tende a excluir mais variáveis, sob risco de perder variáveis que forem importantes para o modelo. O valor "ideal" da mudança mínima dependerá de seus dados e do aplicativo. Consulte o Log de Erro de Seleção de Variável na saída para ajudar a avaliar quais variáveis são mais importantes. Consulte o tópico "Log de erro de seleção de variável" na página 95 para obter mais informações

Partições

A guia Partições permite dividir o conjunto de dados em conjuntos de treinamento e de validação e, quando aplicável, designar casos em dobras de validação cruzada

Partições de Treinamento e de Validação. Este grupo especifica o método de particionamento do conjunto de dados ativo em amostras de treinamento e de validação. A **amostra de treinamento** compreende os registros de dados utilizados para treinar o modelo de vizinho mais próximo, e alguma porcentagem de casos no conjunto de dados deve ser designada para a amostra de treinamento para obter um modelo. A **amostra de validação** é um conjunto independente de registros de dados utilizados para avaliar o modelo final; o erro para a amostra de validação fornece uma estimativa "honesta" da capacidade preditiva do modelo porque os casos de validação não foram utilizados para construir o modelo.

- **Designar aleatoriamente casos às partições.** Especifique a porcentagem de casos para designar à amostra de treinamento. O restante é designado para a amostra de validação.
- **Usar variável para designar casos.** Especifique uma variável numérica que designa a cada caso no conjunto de dados ativo para a amostra de treinamento ou validação. Casos com um valor positivo na variável são designados para a amostra de treinamento, e casos com um valor de 0 ou um valor

negativo para a amostra de validação. Casos com um valor omisso do sistema são excluídos da análise. Quaisquer valores omissos do usuário para a variável de partição são sempre tratados como válidos.

Dobras de Validação Cruzada. A validação cruzada de dobra *V* é utilizada para determinar o "melhor" número de vizinhos. Ela não está disponível junto com a seleção de variável por motivos de desempenho.

A validação cruzada divide a amostra em um número de subamostras ou dobras. Em seguida, os modelos de vizinho mais próximo são gerados, excluindo os dados de cada subamostra por vez. O primeiro modelo baseia-se em todos os casos, exceto aqueles na primeira dobra de amostra, o segundo modelo baseia-se em todos os casos, exceto aqueles na segunda dobra de amostra, e assim por diante. Para cada modelo, o erro é estimado ao aplicar o modelo à subamostra excluída ao gerá-lo. O "melhor" número de vizinhos mais próximos é aquele que produz o menor erro entre as dobras.

- **Designar aleatoriamente casos às dobras.** Especifique o número de dobras que devem ser utilizadas para validação cruzada. O procedimento designa aleatoriamente casos às dobras, numerados de 1 a *V*, que é o número de dobras.
- **Usar variável para designar casos.** Especifique uma variável numérica que designa cada caso no conjunto de dados ativo a uma dobra. A variável deve ser numérica e assumir valores de 1 a *V*. Se quaisquer valores neste intervalo estiverem omissos e, em quaisquer divisões, se os arquivos de divisão estiverem em vigor, isso causará um erro.

Configurar semente para Mersenne Twister. Configurar uma semente permite replicar análises. O uso desse controle é semelhante a configurar o Mersenne Twister como o gerador ativo e especificar um ponto inicial fixo no diálogo Geradores de Número Aleatório, com a diferença importante de que configurar a semente neste diálogo irá preservar o estado atual do gerador de número aleatório e restaurar esse estado após a análise ser concluída.

Salvar

Nomes de Variáveis Salvas. A geração de nome automática assegura que você mantenha todo o seu trabalho. Nomes customizados permitem descartar/substituir resultados de execuções anteriores sem primeiro excluir as variáveis salvas no Editor de Dados.

Variáveis para Salvar

- **Valor predito ou categórico.** Isso salva o valor predito para uma variável resposta escalar ou a categoria predita para uma variável resposta categórica.
- **Probabilidade predita.** Isso salva as probabilidades preditas para uma variável resposta categórica. Uma variável separada é salva para cada uma das *n* primeiras categorias, em que *n* é especificado no controle **Máximo de categorias a serem salvas para variável resposta categórica**.
- **Variáveis de partição de Treinamento/Validação.** Se os casos forem designados aleatoriamente para as amostras de treinamento e de validação na guia Partições, isso salvará o valor da partição (treinamento ou validação) ao qual o caso foi designado.
- **Variável de dobra de validação cruzada.** Se os casos forem designados aleatoriamente para dobras de validação cruzada na guia Partições, isso salvará o valor da dobra à qual o caso foi designado.

Saída

Saída do visualizador

- **Sumarização do processamento de caso.** Exibe a tabela de sumarização de processamento de caso, que sumariza o número de casos incluídos e excluídos na análise, no total e por amostras de treinamento e de validação.
- **Gráficos e tabelas.** Exibe a saída relacionada a modelo, incluindo tabelas e gráficos. As tabelas na visualização de modelo incluem *k* vizinhos mais próximos e distâncias para casos focais, classificação de variáveis resposta categórica e uma sumarização de erros. A saída gráfica na visualização de modelo

inclui um log de erro de seleção, um gráfico de importância de variável, um gráfico de espaço da variável, um gráfico de peers e um mapa de quadrante. Consulte o tópico "Visualização do Modelo" para obter mais informações

Arquivos

- **Exportar modelo em XML.** É possível usar esse arquivo de modelo para aplicar as informações de modelo a outros arquivos de dados para propósitos de escoragem. Essa opção não estará disponível se arquivos de divisão tiverem sido definidos.
- **Exportar distâncias entre casos focais e k vizinhos mais próximos.** Para cada caso focal, uma variável separada é criada para cada um dos *k* vizinhos mais próximos do caso focal (a partir da amostra de treinamento) e para as *k* distâncias mais próximas correspondentes.

Opções

Valores omissos de usuário. Variáveis categóricas devem ter valores válidos para um caso a ser incluído na análise. Estes controles permitem decidir se valores omissos de usuário são tratados como válidos entre variáveis categóricas.

Valores omissos do sistema e valores omissos para variáveis de escala são sempre tratados como inválidos.

Visualização do Modelo

Ao selecionar **Gráficos e tabelas** na guia Saída, o procedimento cria um objeto Modelo de Vizinho Mais Próximo no Visualizador. Ao ativar (clicar duas vezes) nesse objeto, você obtém uma visualização interativa do modelo. A visualização do modelo possui uma janela de 2 painéis:

- O primeiro painel exibe uma visão geral do modelo chamado de visualização principal.
- O segundo painel exibe um dos dois tipos de visualizações:

  Uma visualização do modelo auxiliar mostra mais informações sobre o modelo, mas não está focada no próprio modelo.

  Uma visualização vinculada é uma visualização que mostra detalhes sobre uma variável do modelo quando o usuário realiza drill down na parte da visualização principal.

Por padrão, o primeiro painel mostra o espaço de recursos e o segundo painel mostra o gráfico de importância de variável. Se o gráfico de importância de variável não estiver disponível, ou seja, quando **Ponderar recursos por importância** não foi selecionada na guia Recursos, a primeira visualização disponível no menu suspenso Visualizar será mostrada.

Quando uma visualização não possuir informações disponíveis, seu item de texto no menu suspenso Visualizar será desativado.

Espaço da Variável

O gráfico de espaço da variável é um gráfico interativo do espaço da variável (ou um subespaço, se houver mais de 3 variáveis). Cada eixo representa uma variável no modelo, e a localização dos pontos no gráfico mostra os valores dessas variáveis para casos nas partições de treinamento e de validação.

Chaves. Além dos valores de variável, os pontos no gráfico transmitem outras informações.

- Forma que indica que a partição à qual um ponto pertence, seja Treinamento ou Validação.
- A cor/sombreamento de um ponto indica o valor da resposta para esse caso, com os valores de cores distintos iguais às categorias de uma variável resposta categórica e os sombreamentos indicando o intervalo de valores de uma variável resposta contínua. O valor indicado para a partição de treinamento é o valor observado e, para a partição de validação, é o valor predito. Se nenhuma resposta for especificada, essa chave não será mostrada.

- Estruturas de tópicos mais pesadas indicam que um caso é focal. Casos focais são mostrados vinculados aos seus *k* vizinhos mais próximos.

Controles e Interatividade. Diversos controles no gráfico permitem explorar o Espaço da Variável.

- É possível escolher qual subconjunto de variáveis deseja mostrar no gráfico e alterar quais variáveis são representadas nas dimensões.
- Os "casos focais" são simplesmente pontos selecionados no gráfico de Espaço da Variável. Se você especificou uma variável de caso focal, os pontos que representam os casos focais serão inicialmente selecionados. No entanto, qualquer ponto pode temporariamente se tornar um caso focal se você selecioná-lo. Os controles "comuns" para seleção de ponto se aplicam; clicar em um ponto seleciona esse ponto e desmarca todos os outros; Controle - Clicar em um ponto o inclui no conjunto de pontos selecionados. Visualizações vinculadas, como o Gráfico de Peers, serão atualizadas automaticamente com base nos casos selecionados no Espaço da Variável.
- É possível alterar o número de vizinhos mais próximos (*k*) para exibir para casos focais.
- Passar o mouse sobre um ponto no gráfico exibe uma dica de ferramenta com o valor do rótulo case, ou o número do caso se rótulos case não forem definidos, e os valores de resposta observados e preditos.
- O botão "Reconfigurar" permite retornar o Espaço da Variável para seu estado original.

Incluindo e removendo arquivos/variáveis

É possível incluir novos campos/variáveis no espaço de recursos ou remover aqueles que estão sendo exibidos.

Paleta de variáveis

A paleta de variáveis deve ser exibida antes que seja possível incluir e remover variáveis. Para exibir a paleta de variáveis, o Visualizador de Modelo deve estar no modo Em edição e um caso deve ser selecionado no espaço do recurso.

1. Para colocar o Visualizador de Modelo no modo Em edição, a partir dos menus, escolha:

   Visualizar > **Modo de Edição**
2. Uma vez no modo Em edição, clique em qualquer caso no espaço do recurso.
3. Para exibir a paleta de variáveis, a partir dos menus, escolha:

   Visualizar > **Paletas** > **Variáveis**

   A paleta de variáveis lista todas as variáveis no espaço do recurso. O ícone próximo ao nome da variável indica o nível de medição da variável.
4. Para alterar temporariamente o nível de medição de uma variável, clique com o botão direito do mouse na variável na paleta de variáveis e escolha uma opção.

Zonas de variáveis

As variáveis são incluídas em "zonas" no espaço do recurso. Para exibir as zonas, inicie arrastando uma variável da paleta de variáveis ou selecione **Mostrar zonas**.

O espaço do recurs possui zonas para os eixos *x*, *y* e *z*.

Movendo variáveis nas zonas

Aqui estão algumas regras gerais e dicas para mover as variáveis nas zonas:

- Para mover uma variável em uma zona, clique e arraste a variável da paleta de variáveis e solte-a na zona. Se você escolher **Mostrar zonas**, será possível também clicar com o botão direito do mouse em uma zona e selecione uma variável que você deseja incluir na zona.

- Se você arrastar uma variável a partir da paleta de variáveis para uma zona já ocupada por outra variável, a variável antiga será substituída pelo nova.
- Se você arrastar uma variável de uma zona para uma zona já ocupada por outra variável, as variáveis trocarão de posições.
- Clicar no X em uma zona remove a variável desta zona.
- Se houver vários elementos gráficos na visualização, cada elemento gráfico pode ter sua própria zona de variáveis associadas. Primeiro, selecione o elemento gráfico.

Importância da Variável

Geralmente você desejará focar seus esforços de modelagem nas variáveis que forem mais importantes e considerar descartar ou ignorar aquelas que forem menos importantes. O gráfico de importância de variável ajuda a fazer isso ao indicar a importância relativa de cada variável na estimativa do modelo. Como os valores são relativos, a soma dos valores para todas as variáveis na exibição é 1,0. A importância da variável não se relaciona com a precisão do modelo. Ela está relacionada somente com a importância de cada variável em fazer uma predição, e não se a predição é precisa ou não.

Peers

Este gráfico exibe os casos focais e seus *k* vizinhos mais próximos em cada recurso e na resposta. Ele estará disponível se um caso focal for selecionado no Espaço de Recursos.

Vinculando comportamento. O gráfico Peers é vinculado ao Espaço da Variável de duas maneiras.

- Casos selecionados (focais) no Espaço da Variável são exibidos no gráfico de Peers, junto de seus *k* vizinhos mais próximos.
- O valor de *k* selecionado no Espaço Recurso é usado no gráfico de Peers.

Distâncias de Vizinho Mais Próximo

Esta tabela exibe os *k* vizinhos mais próximos e as distâncias somente para casos focais. Ela estará disponível se um identificador de caso focal for especificado na guia Variáveis, e exibe somente os casos focais identificados por essa variável.

Cada linha da:

- A coluna **Caso Focal** contém o valor da variável de rótulo case para o caso focal; se rótulos case não forem definidos, essa coluna conterá o número do caso do caso focal.
- A *i*-ésima coluna no grupo Vizinhos Mais Próximos contém o valor da variável de rotulagem de caso para o *i*-ésimo vizinho mais próximo do caso focal; se rótulos case não forem definidos, esta coluna conterá o número do caso do *i*-ésimo vizinho mais próximo do caso focal.
- A *i*-ésima coluna no grupo Distâncias Mais Próximas contém a distância do *i*-ésimo vizinho mais próximo do caso focal

Mapa de Quadrante

Este gráfico exibe os casos focais e seus *k* vizinhos mais próximos em um gráfico de dispersão (ou gráfico de pontos, dependendo do nível de medição da resposta) com a resposta no eixo *y* e uma variável de escala no eixo *x*, agrupados em painéis por variáveis. Ele estará disponível se houver uma resposta e se um caso focal estiver selecionado no Espaço de Variável.

- As linhas de referência são desenhadas para variáveis contínuas, nas médias da variável na partição de treinamento.

Log de erro de seleção de variável

Pontos no gráfico exibem o erro (a taxa de erros ou a soma dos quadrados dos erro, dependendo do nível de medição da resposta) no eixo *y* para o modelo com o recurso listado no eixo *x* (além de todos os recursos à esquerda no eixo *x*). Este gráfico estará disponível se houver uma seleção de resposta e de variável em vigor.

Log de erro de seleção k

Pontos no gráfico exibem o erro (a taxa de erros ou a soma dos quadrados dos erros, dependendo do nível de medição da resposta) no eixo *y* para o modelo com o número de vizinhos mais próximos (*k*) no eixo *x*. Este gráfico estará disponível se houver uma resposta e uma seleção *k* estiver em vigor.

Log de Erro de k e Seleção de variável

Esses são gráficos de seleção de variável (consulte "Log de erro de seleção de variável"), agrupados em painéis por *k*. Este gráfico estará disponível se houver uma resposta e se *k* e uma seleção de variável estiverem em vigor.

Tabela de Classificação

Esta tabela exibe a classificação cruzada de valores observados versus valores preditos da resposta, por partição. Ela estará disponível se houver uma resposta e for categórica.

- A linha **(Omisso)** na partição Validação contém casos de validação com valores omissos na resposta. Estes casos contribuem com a Amostra de Validação: Valores de Porcentagem Geral, mas não para valores de Porcentagem Correta.

Sumarização de Erro

Esta tabela estará disponível se houver uma variável de destino. Ela exibe o erro associado ao modelo; a soma dos quadrados para uma variável resposta contínua e a taxa de erro (100% – percentual geral correto) para uma variável resposta categórica.

Capítulo 21. Análise discriminante

A análise discriminante constrói um modelo preditivo para associação ao grupo. O modelo é composto de uma função discriminante (ou, para mais de dois grupos, um conjunto de funções discriminantes) com base nas combinações lineares das variáveis preditoras que fornecem a melhor discriminação entre os grupos. As funções são geradas a partir de uma amostra de casos para os quais a associação ao grupo é conhecida e podem, em seguida, ser aplicadas aos novos casos que tiverem medições para as variáveis preditoras e possuírem associação ao grupo desconhecida.

Nota: A variável de agrupamento pode ter mais de dois valores. Os códigos para a variável de agrupamento devem ser números inteiros, no entanto, é necessário especificar seus valores mínimo e máximo. Casos com valores fora desses limites são excluídos da análise.

Exemplo. Em média, pessoas que vivem em países de zona temperada consumem mais calorias por dia do que pessoas que vivem em zonas tropicais, e uma maior proporção de pessoas nas zonas temperadas representa habitantes das cidades. Um pesquisador deseja combinar essas informações em uma função para determinar quão bem um indivíduo pode diferenciar-se entre os dois grupos de países. O pesquisador acredita que o tamanho da população e informações econômicas também podem ser importantes. A análise discriminante permite estimar os coeficientes da função discriminante linear, que é semelhante ao lado direito de uma equação de regressão linear múltipla. Ou seja, utilizando os coeficientes *a*, *b*, *c* e *d*, a função é:

```
D = a * climate + b * urban + c * population + d * gross domestic product per capita
```

Se essas variáveis forem úteis para discriminar entre as duas zonas climáticas, os valores de *D* irão diferir para países de zonas temperadas e tropicais. Se utilizar um método de seleção de variáveis stepwise, talvez você ache que não é necessário incluir todas as quatro variáveis na função.

Estatísticas. Para cada variável: médias, desvios padrão, ANOVA univariada. Para cada análise: *M* de Box, matriz de correlações dentro de grupos, matriz de covariâncias dentro de grupos, matriz de covariâncias de grupos separados, matriz de covariâncias totais. Para cada função discriminante canônica: autovalor, porcentagem de variância, correlação canônica, Lambda de Wilks e qui-quadrado. Para cada passo: probabilidades anteriores, coeficientes de função de Fisher, coeficientes função não padronizados, Lambda de Wilks' para cada função canônica.

Considerações de Dados de Análise Discriminante

Dados. A variável de agrupamento deve ter um número limitado de categorias distintas, codificadas como números inteiros. As variáveis independentes que forem nominais devem ser recodificadas em variáveis simuladas ou de contraste.

Suposições. Os casos devem ser independentes. As variáveis preditoras devem ter uma distribuição normal multivariada, e matrizes de variância-covariância dentro de grupo devem ser iguais entre os grupos. A associação ao grupo é considerada mutuamente exclusiva (ou seja, nenhum caso pertence a mais de um grupo) e coletivamente exaustiva (ou seja, todos os casos são membros de um grupo). O procedimento é mais eficiente quando a associação ao grupo for uma variável realmente categórica; se a associação ao grupo for baseada em valores de uma variável contínua (por exemplo, QI alto e QI baixo), considere utilizar regressão linear para aproveitar as informações mais completas que forem oferecidas pela própria variável contínua.

Para Obter uma Análise Discriminante

1. Nos menus, escolha:

   Analisar > **Classificar** > **Discriminante...**

2. Selecione uma variável de agrupamento com valor de número inteiro e clique em **Definir Intervalo** para especificar as categorias de interesse.
3. Selecione as variáveis independentes ou preditoras. (Se a sua variável de agrupamento não possuir valores de número inteiro, o menu Recodificação Automática na Transformação criará uma variável que possua).
4. Selecione o método para inserir as variáveis independentes.
   - **Inserir independentes juntas.** Insere simultaneamente todas as variáveis independentes que satisfizerem os critérios de tolerância.
   - **Utilizar o método stepwise.** Usa análise stepwise para controlar entrada e remoção de variável.
5. Opcionalmente, selecione os casos com uma variável de seleção.

Análise Discriminante para Definir Intervalo

Especifique o valor mínimo e máximo da variável de agrupamento para a análise. Casos com valores fora deste intervalo não são utilizados na análise discriminante, mas são classificados em um dos grupos existentes com base nos resultados da análise. Os valores mínimo e máximo devem ser números inteiros.

Selecionar Casos para Análise Discriminante

Para selecionar casos para sua análise:

1. Na caixa de diálogo Análise Discriminante, escolha uma variável de seleção.
2. Clique em **Valor** para inserir um número inteiro como o valor de seleção.

Apenas casos com o valor especificado para a variável de seleção são utilizados para derivar as funções discriminantes. Estatísticas e resultados de classificação são gerados para casos selecionados e não selecionados. Esse processo fornece um mecanismo para classificar casos novos com base em dados existentes anteriormente ou para particionar seus dados em subconjuntos de treinamento e de teste para executar a validação no modelo gerado.

Estatística de Análise Discriminante

Descritivos. As opções disponíveis são médias (incluindo desvios padrão), ANOVAs univariados e teste *M* de Box.

- *Médias*. Exibe as médias totais e de grupo, bem como os desvios padrão para as variáveis independentes.
- *ANOVAs Univariadas*. Executa um teste de análise de variância unidirecional de igualdade de médias de grupo para cada variável independente.
- *M de Box*. Um teste para a igualdade das matrizes de covariâncias de grupo. Para amostras suficientemente grandes, um valor p não significativo representa que não há evidência suficiente de que as matrizes diferem. O teste é sensível a partidas da normalidade multivariada.

Coeficientes de Função. As opções disponíveis são coeficientes de classificação de Fisher e coeficientes não padronizados.

- *de Fisher*. Exibe os coeficientes da função de classificação de Fisher que podem ser utilizados diretamente para classificação. Um conjunto separado de coeficientes de função de classificação é obtido para cada grupo, e um caso é designado ao grupo para o qual ele possui o maior escore discriminante (valor da função de classificação).
- *Não padronizado*. Exibe os coeficientes de função discriminante não padronizadas.

Matrizes. As matrizes de coeficientes disponíveis para variáveis independentes são matriz de correlações dentro de grupos, matriz de covariâncias dentro de grupos, matriz de covariâncias separada de grupos e o total de matrizes de covariância.

- *Correlação dentro de grupos*. Exibe uma matriz de correlações dentro de grupos em conjunto que é obtida pela média das matrizes de covariâncias separadas de todos os grupos antes de calcular as correlações.
- *Covariância dentro de grupos*. Exibe uma matriz de covariâncias dentro de grupos em conjunto, que pode diferir da matriz de covariâncias totais. A matriz é obtida pela média das matrizes covariâncias separadas para todos os grupos.
- *Covariância separada de grupos*. Exibe matrizes de covariância separadas para cada grupo.
- *Total de covariâncias*. Exibe uma matriz de covariâncias a partir de todos os casos como se fossem de uma única amostra.

Método Stepwise para Análise Discriminante

Método. Selecione a estatística a ser utilizada para inserir ou remover novas variáveis. As alternativas disponíveis são lambda de Wilks, variância não explicada, distância de Mahalanobis, razão de *F* menor e *V* de Rao. Com o *V* de Rao, é possível especificar o aumento mínimo em *V* para uma variável a ser inserida.

- *Lambda de Wilks*. Um método de seleção de variáveis para análise discriminante stepwise que escolhe variáveis a serem inseridas na equação com base no quanto elas diminuem o lambda de Wilks. Em cada passo, a variável que minimiza o lambda geral de Wilks é inserida.
- *Variância não explicada*. Em cada passo, a variável que minimiza a soma da variação não explicada entre os grupos é inserida.
- *Distância de Mahalanobis*. A medida do quanto os valores de um caso nas variáveis independentes diferem da média de todos os casos. Uma distância de Mahalanobis grande identifica um caso como tendo valores extremos em uma ou mais variáveis independentes.
- *Razão F menor*. Um método de seleção de variáveis na análise stepwise com base na maximização de uma razão F calculada a partir da distância de Mahalanobis entre os grupos.
- *V de Rao*. A medida das diferenças entre as médias de grupo. Também chamada de rastreio de Lawley-Hotelling. Em cada passo, a variável que maximiza o aumento no V de Rao é inserida. Após selecionar esta opção, insira o valor mínimo que uma variável deve ter para inserir na análise.

Critérios. As alternativas disponíveis são **Usar valor F** e **Usar probabilidade de F**. Insira valores para inserção e remoção de variáveis.

- *Usar valor F*. Uma variável será inserida no modelo se seu valor F for maior que o valor de Entrada e será removida se o valor F for menor que o valor de Remoção. A Entrada deve ser maior que Remoção, e ambos os valores devem ser positivos. Para inserir mais variáveis no modelo, diminua o valor de Entrada. Para remover mais variáveis do modelo, aumente o valor de Remoção.
- *Usar probabilidade de F*. Uma variável será inserida no modelo se o nível de significância de seu valor F for menor que o valor de Entrada e será removida se o nível de significância for maior que o valor de Remoção. A Entrada deve ser menor que Remoção, e ambos os valores devem ser positivos. Para inserir mais variáveis no modelo, aumente o valor de Entrada. Para remover mais variáveis do modelo, diminua o valor de Remoção.

Exibição. Sumarização dos passos exibe as estatísticas para todas as variáveis após cada passo; **F para distâncias entre pares** exibe uma matriz de razões de *F* entre pares para cada par de grupos.

Classificação para Análise Discriminante

Probabilidades Anteriores. Esta opção determina se os coeficientes de classificação são ajustados para um conhecimento a priori de associação ao grupo.

- **Todos os grupos iguais.** As probabilidades anteriores iguais são assumidas para todos os grupos; isso não tem efeito sobre os coeficientes.
- **Calcular a partir de tamanhos de grupo.** Os tamanhos de grupo observados em sua amostra determinam as probabilidades anteriores de associação ao grupo. Por exemplo, se 50% das observações

incluídas na análise caírem no primeiro grupo, 25% no segundo e 25% no terceiro, os coeficientes de classificação serão ajustados para aumentar a probabilidade de associação no primeiro grupo relativo aos outros dois.

Exibição. As opções de exibição disponíveis são resultados entre casos, tabela de sumarização e classificação com exclusão de um item.

- *Resultados do caso*. Os códigos para o grupo real, o grupo predito, as probabilidades posteriores e os escores discriminantes são exibidos para cada caso.
- *Tabela de sumarização*. O número de casos designados correta e incorretamente para cada um dos grupos com base na análise discriminante. Às vezes é chamado de "Matriz de Confusão".
- *Classificação com exclusão de um item*. Cada caso na análise é classificado pelas funções derivadas de todos os outros casos diferentes desse caso. Também é conhecida como "Método U".

Substituir valores omissos pela média. Selecione esta opção para substituir a média de uma variável independente para um valor omissos somente durante a fase de classificação.

Usar Matriz de Covariâncias. É possível optar por classificar casos utilizando uma matriz de covariâncias dentro de grupos ou uma matriz de covariâncias separada de grupos.

- *Dentro de grupos*. A matriz de covariâncias dentro de grupos em conjunto é usada para classificar casos.
- *Separado de grupos*. As matrizes de covariância grupos-separados são usadas para classificação. Como a classificação é baseada nas funções discriminantes (não com base nas variáveis originais), essa opção nem sempre é equivalente à discriminação quadrática.

Gráficos. As opções de gráfico disponíveis são grupos combinados, grupos separados e mapa territorial.

- *Grupos combinados*. Cria um gráfico de dispersão de todos os grupos dos dois primeiros valores da função discriminante. Se houver apenas uma função, um histograma será exibido.
- *Separado de grupos*. Cria gráficos de dispersão de grupos separados dos dois primeiros valores da função discriminante. Se houver apenas uma função, histogramas serão exibidos.
- *Mapa territorial*. Um gráfico dos limites utilizados para classificar os casos em grupos com base nos valores de função. Os números correspondem aos grupos nos quais os casos são classificados. A média de cada grupo é indicada por um asterisco dentro de seus limites. O mapa não será exibido se houver apenas uma função discriminante.

Salvamento de Análise Discriminante

É possível incluir novas variáveis em seu arquivo de dados ativos. As opções disponíveis são associação ao grupo predito (uma variável única), escores discriminantes (uma variável para cada função discriminante na solução), e probabilidades de associação ao grupo, dados os escores discriminantes (uma variável para cada grupo).

Também é possível exportar informações de modelo para o arquivo especificado no formato XML. É possível usar esse arquivo de modelo para aplicar as informações de modelo a outros arquivos de dados para propósitos de escoragem.

Recursos Adicionais do Comando DISCRIMINANT

O idioma da sintaxe de comando também permite:

- Executar diversas análises discriminante (com um comando) e controlar a ordem na qual as variáveis são inseridas (com o subcomando `ANALYSIS`).
- Especificar as probabilidades anteriores para classificação (com o subcomando `PRIORS`).
- Exibir padrões e matrizes de estrutura girados (com o subcomando `ROTATE`).
- Limitar o número de funções discriminantes extraídas (com o subcomando `FUNCTIONS`).

- Restringir classificação para os casos que forem selecionados (ou desmarcados) para a análise (com o subcomando SELECT).
- Ler e analisar uma matriz de correlações (com o subcomando MATRIX).
- Gravar uma matriz de correlações para análise posterior (com o subcomando MATRIX).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 22. Análise fatorial

A análise fatorial tenta identificar variáveis subjacentes, ou **fatores**, que explicam o padrão de correlações dentro de um conjunto de variáveis observadas. A análise fatorial é geralmente utilizada na redução de dados para identificar um número pequeno de fatores que explicam a maior parte da variância que é observada em um número muito maior de variáveis de manifesto. A análise fatorial também pode ser utilizada para gerar hipóteses referentes a mecanismos causais ou para verificar variáveis para análise subsequente (por exemplo, para identificar colinearidade antes de executar uma análise de regressão linear).

O procedimento de análise fatorial oferece um alto grau de flexibilidade:

- Sete métodos de extração de fator estão disponíveis.
- Cinco métodos de rotação estão disponíveis, incluindo oblimin e promax diretos para rotações não ortogonais.
- Três métodos de calcular os escores dos fatores estão disponíveis, e os escores podem ser salvos como variáveis para análise adicional.

Exemplo. Que atitudes subjacentes levam as pessoas a responderem perguntas em uma pesquisa de opinião política da maneira que fazem? Examinar as correlações entre os itens de pesquisa de opinião revela que há uma sobreposição significativa entre os vários subgrupos de itens - perguntas sobre impostos tendem a se correlacionar entre si, perguntas sobre assuntos militares se correlacionam entre si, e assim por diante. Com a análise fatorial, é possível investigar o número de fatores subjacentes e, em muitos casos, identificar o que os fatores representam conceitualmente. Além disso, é possível calcular os escores dos fatores de cada respondente, que podem ser utilizados em análises subsequentes. Por exemplo, é possível construir um modelo de regressão logística para prever o comportamento de voto com base nos escores dos fatores.

Estatísticas. Para cada variável: número de casos válidos, média e desvio padrão. Para cada análise de fator: matriz de correlações de variáveis, incluindo níveis de significância, determinante, e o inverso; matriz de correlações reproduzidas, incluindo anti-imagem; solução inicial (comunalidades, autovalores e a porcentagem da variância explicada); medida da adequação de amostragem de Kaiser-Meyer-Olkin e teste de esfericidade de Bartlett; solução não girada, incluindo carregamentos de fatores, comunalidades e autovalores; e solução girada, incluindo a matriz de padrões girada e a matriz de transformação. Para rotações oblíquas: matrizes de padrão e de estrutura girada, matriz de coeficiente de escore dos fatores e matriz de covariâncias de fatores. Gráficos: scree plot de autovalores e gráfico de carregamento dos dois ou três primeiros fatores.

Considerações de Dados de Análise Fatorial

Dados. As variáveis devem ser quantitativas no nível de *intervalo* ou *razão*. Dados categóricos (como religião ou país de origem) não são adequados para análise fatorial. Os dados para os quais os coeficientes de correlação de Pearson podem ser calculados devem ser adequados para análise fatorial.

Suposições. Os dados devem ter uma distribuição normal bivariada para cada par de variáveis, e as observações devem ser independentes. O modelo de análise fatorial especifica que as variáveis são determinadas por fatores comuns (os fatores estimados pelo modelo) e por fatores exclusivos (que não se sobrepõem entre as variáveis observadas), e as estimativas calculadas baseiam-se na suposição de que todos os fatores exclusivos não são correlacionados uns com os outros e com os fatores comuns.

Para Obter uma Análise Fatorial

1. Nos menus, escolha:

   Analisar > **Redução de Dimensão** > **Fator...**

2. Selecione as variáveis para a análise fatorial.

Selecionar Casos para Análise Fatorial

Para selecionar casos para sua análise:

1. Escolha uma variável de seleção.
2. Clique em **Valor** para inserir um número inteiro como o valor de seleção.

Apenas casos com esse valor para a variável de seleção são usados na análise fatorial.

Descritivas de Análise Fatorial

Estatísticas. Descritivas univariadas inclui a média, o desvio padrão e o número de casos válidos para cada variável. **Solução inicial** exibe comunalidades iniciais, autovalores e a porcentagem da variância explicada.

Matriz de Correlações. As opções disponíveis são coeficientes, níveis de significância, determinante, KMO e teste de esfericidade de Bartlett, inverso, reproduzido e anti-imagem.

- *KMO e Teste de Esfericidade de Bartlett (Análise Fatorial)* A medida de adequação de amostragem de Kaiser-Meyer-Olkin testa se as correlações parciais entre as variáveis são pequenas. O teste de esfericidade de Bartlett testa se a matriz de correlações é uma matriz de identidade, que indicaria que o modelo de fator é inadequado.
- *Reproduzido*. A matriz de correlações estimadas a partir da solução de fator. Os resíduos (diferença entre correlações estimadas e observadas) também são exibidos.
- *Anti-imagem*. A matriz de correlações anti-imagem contém os negativos dos coeficientes de correlação parciais e a matriz de covariâncias anti-imagem contém os negativos das covariâncias parciais. Em um fator de modelo bom, muitos dos elementos fora da diagonal serão pequenos. A medida da adequação de amostragem para uma variável é exibida na diagonal da matriz de correlações anti-imagem.

Extração de Análise Fatorial

Método. Permite especificar o método de extração de fator. Os métodos disponíveis são componentes principais, quadrados mínimos não ponderados, quadrados mínimos generalizados, máxima verossimilhança, fatoração de eixo principal, fatoração alfa e fatoração de imagem.

- *Análise de Componentes Principais*. Um método de extração de fator utilizado para formar combinações lineares não correlacionadas das variáveis observadas. O primeiro componente possui variância máxima. Componentes sucessivos explicam progressivamente partes menores da variância e nenhum deles se correlaciona com o outro. A análise de componentes principais é utilizada para obter a solução de fator inicial. Ela pode ser usada quando uma matriz de correlações é singular.
- *Método de Quadrados Mínimos Não Ponderados*. Um método de extração de fator que minimiza a soma das diferenças quadráticas entre as matrizes de correlação observadas e reproduzidas (ignorando as diagonais).
- *Método de Quadrados Mínimos Generalizado*. Um método de extração de fator que minimiza a soma das diferenças quadráticas entre as matrizes de correlação observadas e reproduzidas. As correlações são ponderadas pelo inverso de sua exclusividade, para que as variáveis com exclusividade alta recebam uma ponderação menor que aquelas com exclusividade baixa.
- *Método de Máxima Verossimilhança*. Um método de extração de fator que produz estimativas paramétrica que mais provavelmente produziram a matriz de correlações observadas se a amostra for de uma distribuição normal multivariada. As correlações são ponderadas pelo inverso da exclusividade das variáveis, e um algoritmo iterativo é utilizado.
- *Método de eixo principal de análise fatorial*. Um método de extrair fatores da matriz de correlações original, com coeficientes de correlação múltipla quadrada colocados na diagonal como estimativas iniciais de comunalidades. Esses carregamentos de fatores são utilizados para estimar novas

comunalidades que substituem as estimativas de comunalidades antigas na diagonal. As iterações continuam até que as mudanças nas comunalidades de uma iteração para a próxima satisfaçam o critério de convergência para a extração.

- *Alpha*. Um método de extração de fator que considera as variáveis na análise uma amostra do universo de possíveis variáveis. Esse método maximiza a confiabilidade alfa dos fatores.
- *Método de imagem de análise fatorial*. Um método de extração de fator desenvolvido por Guttman e com base em teoria de imagem. A parte comum da variável, chamada de imagem parcial, é definida como sua regressão linear em variáveis restantes, ao invés de uma função de fatores hipotéticos.

Analisar. Permite especificar uma matriz de correlações ou uma matriz de covariâncias.

- **Matriz de correlações.** Útil se variáveis em sua análise forem medidas em escalas diferentes.
- **Matriz de covariâncias.** Útil quando desejar aplicar sua análise fatorial a diversos grupos com diferentes variâncias para cada variável.

Extrair. É possível reter todos os fatores cujos autovalores excederem um valor especificado ou reter um número específico de fatores.

Exibição. Permite solicitar a solução de fator não girada e um scree plot dos autovalores.

- *Solução de Fator Não Girado*. Exibe carregamentos de fatores não girados (matriz padrão de fator), comunalidades e os autovalores para a solução de fator.
- *Scree plot*. Um gráfico de variância que é associado a cada fator. Esse gráfico é utilizado para determinar quantos fatores devem ser mantidos. Normalmente, o gráfico mostra uma divisão distinta entre a inclinação íngreme dos fatores grandes e o acompanhamento gradual do restante (o scree).

Máximo de Iterações para Convergência. Permite especificar o número máximo de passos que o algoritmo pode utilizar para estimar a solução.

Rotação de Análise Fatorial

Método. Permite selecionar o método de rotação de fator. Os métodos disponíveis são varimax, oblimin direto, quartimax, equamax ou promax.

- *Método Varimax*. Um método de rotação ortogonal que minimiza o número de variáveis que possuem altos carregamentos em cada fator. Este método simplifica a interpretação dos fatores.
- *Método Oblimin Direto* Um método para rotação oblíqua (não ortogonal). Quando delta é igual a 0 (o padrão), as soluções são mais oblíquas. Conforme delta se torna mais negativo, os fatores se tornam menos oblíquos. Para substituir o delta padrão de 0, insira um número menor ou igual a 0,8.
- *Método Quartimax*. Um método de rotação que minimiza o número de fatores necessários para explicar cada variável. Este método simplifica a interpretação das variáveis observadas.
- *Método Equamax*. Um método de rotação é uma combinação do método varimax, que simplifica os fatores, e do método quartimax, que simplifica as variáveis. O número de variáveis que são altamente carregadas em um fator e o número de fatores necessários para explicar uma variável são minimizados.
- *Rotação Promax*. Uma rotação oblíqua, que permite que os fatores sejam correlacionados. Essa rotação pode ser calculada mais rapidamente do que uma rotação oblimin direta, sendo útil para conjuntos de dados grandes.

Exibição. Permite incluir a saída na solução girada, bem como carregar gráficos para os primeiros dois ou três fatores.

- *Solução Girada*. Um método de rotação deve ser selecionado para obter uma solução girada. Para rotações ortogonais, a matriz de padrões girada e a matriz de transformação de fator são exibidas. Para rotações oblíquas, as matrizes de correlação de padrão, de estrutura e de fator são exibidas.
- *Gráfico de Carregamento de Fator*. Gráfico de carregamento de fator tridimensional dos três primeiros fatores. Para uma solução de dois fatores, um gráfico bidimensional é mostrado. O gráfico não será exibido se apenas um fator for extraído. Os gráficos exibem soluções giradas se a rotação for solicitada.

Máximo de Iterações para Convergência. Permite especificar o número máximo de passos que o algoritmo pode utilizar para executar a rotação.

Escores de Análise Fatorial

Salvar como variáveis. Cria uma nova variável para cada fator na solução final.

Método. Os métodos alternativos para calcular os escores dos fatores são de regressão, Bartlett e Anderson-Rubin.

- *Método de regressão*. Um método para estimar os coeficientes de escore dos fatores. Os escores que são produzidos têm uma média de 0 e uma variância igual à correlação múltipla quadrada entre os escores dos fatores estimados e os valores de fatores reais. Os escores podem ser correlacionados mesmo quando fatores forem ortogonais.
- *Escores de Bartlett*. Um método de estimar os coeficientes de escore dos fatores. Os escores que são produzidos possuem uma média de 0. A soma dos quadrados dos fatores exclusivos sobre o intervalo das variáveis é minimizada.
- *Método de Anderson-Rubin*. Um método de estimar os coeficientes de escore dos fatores; uma modificação do método de Bartlett que assegura ortogonalidade dos fatores estimados. Os escores que são produzidos possuem uma média de 0, um desvio padrão de 1 e são não correlacionados.

Exibir matriz do coeficiente de escore dos fatores. Mostra os coeficientes pelos quais as variáveis são multiplicadas para obter os escores dos fatores. Também mostra as correlações entre os escores dos fatores.

Opções de Análise Fatorial

Valores omissos. Permite especificar como os valores omissos são manipulados. As opções disponíveis são excluir casos *listwise*, excluir casos *entre pares* ou substituir pela média.

Formato de Exibição de Coeficiente. Permite controlar os aspectos das matrizes de saída. É possível ordenar os coeficientes pelo tamanho e suprimir coeficientes com valores absolutos que forem menores que o valor especificado.

Recursos Adicionais do Comando FACTOR

O idioma da sintaxe de comando também permite:

- Especificar critérios de convergência para a iteração durante a extração e rotação.
- Especificar gráficos de fatores girados individuais.
- Especificar quantos escores de fatores devem ser salvos.
- Especificar valores diagonais para o método de fatoração do eixo principal.
- Gravar matrizes de correlações ou matrizes de Carregamentos Fatoriais no disco para análise posterior.
- Ler e analisar as matrizes de correlações ou matrizes de Carregamentos Fatoriais.

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 23. Escolhendo um Procedimento para Clusterização

Análises de cluster podem ser executadas utilizando o procedimento de Análise de Cluster TwoStep, Hierárquica ou por K-Médias. Cada procedimento usa um algoritmo diferente para criar clusters, e cada um possui opções não disponíveis nos outros.

Análise de Cluster TwoStep. Para muitos aplicativos, o procedimento de Análise de Cluster TwoStep será o método de escolha. Ele fornece os seguintes recursos exclusivos:

- Seleção automática do melhor número de clusters, além de medidas de escolha entre os modelos de cluster.
- Capacidade de criar modelos de cluster simultaneamente com base em variáveis categóricas e contínuas.
- Capacidade de salvar o modelo de cluster em um arquivo XML externo e, em seguida, ler esse arquivo e atualizar o modelo de cluster utilizando dados mais recentes.

Além disso, o procedimento de Análise de Cluster TwoStep pode analisar arquivos de dados grandes.

Análise de Cluster Hierárquica. O procedimento de Análise de Cluster Hierárquica é limitado a arquivos de dados menores (centenas de objetos a serem armazenados em cluster) e possui os seguintes recursos exclusivos:

- Capacidade para armazenar em cluster casos ou variáveis.
- Capacidade para calcular um intervalo de soluções possíveis e salvar as associações de cluster para cada uma dessas soluções.
- Vários métodos para formação do cluster, transformação de variável e medir a dissimilaridade entre os clusters.

Contanto que todas as variáveis sejam do mesmo tipo, o procedimento de Análise de Cluster Hierárquica pode analisar variáveis de intervalo (contínuas), de contagem ou binárias.

Análise de Cluster por K-Médias. O procedimento Análise de Cluster por K-Médias é limitado a dados contínuos, requer a especificação do número de cluster antecipadamente e possui os seguintes recursos exclusivos:

- Capacidade para salvar distâncias de centros do cluster para cada objeto.
- Capacidade de ler centros do cluster iniciais e de salvar centros do cluster finais em um arquivo do IBM SPSS Statistics externo.

Além disso, o procedimento de Análise de Cluster por K-Médias pode analisar arquivos de dados grandes.

Capítulo 24. Análise de cluster de duas etapas

O procedimento Análise de Cluster TwoStep é uma ferramenta exploratória projetada para revelar agrupamentos naturais (ou clusters) dentro de um conjunto de dados que de outra forma não seriam aparentes. O algoritmo utilizado por este procedimento possui vários recursos desejáveis que o diferenciam das técnicas tradicionais de clusterização:

- **Manipulação de variáveis categóricas e contínuas.** Ao considerar variáveis a serem independentes, uma distribuição multinomial normal conjunta pode ser colocada em variáveis categóricas e contínuas.
- **Seleção automática do número de clusters.** Ao comparar os valores de um critério de escolha de modelo em diferentes soluções de clusterização, o procedimento poderá determinar automaticamente o número ideal de clusters.
- **Escalabilidade.** Ao construir uma árvore de recursos do cluster (CF) que sumariza os registros, o algoritmo TwoStep permite analisar arquivos de dados grandes.

Exemplo. Empresas de produtos de varejo e de consumidor aplicam regularmente técnicas de clusterização aos dados que descrevem os hábitos de compras, sexo, idade, nível de renda, etc. de seus clientes. Essas empresas customizam suas estratégias de marketing e de desenvolvimento de produto para cada grupo de clientes para aumentar as vendas e construir fidelidade à marca.

Medida de Distância. Esta seleção determina como a similaridade entre dois clusters é calculada.

- **Log da verossimilhança.** A medida de probabilidade coloca uma distribuição de probabilidade nas variáveis. Variáveis contínuas são consideradas como sendo distribuídas normalmente, ao passo que as variáveis categóricas são consideradas como sendo multinomiais. Todas as variáveis são consideradas independentes.
- **Euclidiana.** A medida Euclidiana é a distância em "linha reta" entre dois clusters. Ela pode ser utilizada somente quando todas as variáveis forem contínuas.

Número de Clusters. Essa seleção permite especificar como o número de clusters deve ser determinado.

- **Determinar automaticamente.** O procedimento irá determinar automaticamente o "melhor" número de clusters, utilizando os critérios especificados no grupo de Critério de Armazenamento em Cluster. Opcionalmente, insira um número inteiro positivo que especifica o número máximo de clusters que o procedimento deve considerar.
- **Especificar fixo.** Permite corrigir o número de clusters na solução. Insira um número inteiro positivo.

Contagem de Variáveis Contínuas. Este grupo fornece uma sumarização das especificações de padronização de variável contínua feitas na caixa de diálogo Opções. Consulte o tópico “Opções de Análise de Cluster TwoStep” na página 110 para obter mais informações

Critério de clusterização. Esta seleção determina como o algoritmo de clusterização automático determina o número de clusters. O Critério de Informação Bayesiano (BIC) ou o Akaike Information Criterion (AIC) podem ser especificados.

Considerações de Dados de Análise de Cluster TwoStep

Dados. Este procedimento funciona com variáveis contínuas e categóricas. Os casos representam objetos a serem armazenados em cluster, e as variáveis representam atributos nos quais o clusterização é baseado.

Ordem de Caso. Observe que a árvore de recursos do cluster e a solução final podem depender da ordem de casos. Para minimizar os efeitos da ordem, ordene aleatoriamente os casos. Talvez você queira obter várias soluções diferentes com casos classificados em diferentes ordens aleatórias para verificar a

estabilidade de uma determinada solução. Em situações em que isso é difícil devido a tamanhos de arquivos extremamente grandes, diversas execuções com uma amostra de casos classificados em diferentes ordens aleatórias podem ser substituídas.

Suposições. A medida de distância de probabilidade supõe que as variáveis no modelo de cluster são independentes. Além disso, cada variável contínua é considerada como tendo uma distribuição normal (gaussiana), e cada variável categórica é considerada como tendo uma distribuição multinomial. Teste interno empírico indica que o procedimento é bastante robusto para violações da suposição de independências e de suposições de distribuição, mas é necessário reconhecer o quão bem essas suposições são atendidas.

Utilize o procedimento Correlações Bivariadas para testar a independência das duas variáveis contínuas. Utilize o procedimento Crosstabs para testar a independência das duas variáveis categóricas. Utilize o procedimento Médias para testar a independência entre uma variável contínua e uma variável categórica. Utilize o procedimento Explorar para testar a normalidade de uma variável contínua. Utilize o procedimento Teste Qui-Quadrado para testar se uma variável categórica possui uma distribuição multinomial especificada.

Para Obter uma Análise de Cluster TwoStep

1. Nos menus, escolha:

   Analisar > **Classificar** > **Cluster TwoStep...**
2. Selecione uma ou mais variáveis categóricas ou contínuas.

Como opção, você pode:

- Ajustar os critérios pelos quais clusters são construídos.
- Selecionar as configurações para tratamento de ruído, alocação de memória, padronização de variável e entrada de modelo de cluster.
- Solicitar a saída do visualizador de modelo.
- Salvar resultados do modelo no arquivo de trabalho ou em um arquivo XML externo.

Opções de Análise de Cluster TwoStep

Tratamento de Valor Discrepante. Este grupo permite tratar valores discrepantes especialmente durante clusterização se a árvore de recursos de cluster (CF) for preenchida. A árvore CF estará cheia se ela não puder aceitar mais nenhum caso em um nó folha e nenhum nó folha puder ser dividido.

- Se selecionar tratamento de ruído e a árvore CF for preenchida, a árvore crescerá novamente após colocar os casos nas folhas esparsas em uma folha de "ruído". Uma folha será considerada esparsa se ela contiver menos do que a porcentagem especificada de casos do tamanho máximo de folha. Após a árvore ter crescido novamente, os valores discrepantes serão colocados na árvore CF, se possível. Caso contrário, os valores discrepantes serão descartados.
- Se você não selecionar tratamento de ruído e a árvore CF for preenchida, ela crescerá novamente utilizando um limite de mudança de distância maior. Após o clusterização final, os valores que não puderem ser designados a um cluster serão rotulados como valores discrepantes. O cluster de valor discrepante recebe um número de identificação de -1 e não é incluído na contagem do número de clusters.

Alocação de Memória. Este grupo permite especificar a quantia máxima de memória em megabytes (MB) que o algoritmo de cluster deve utilizar. Se o procedimento exceder esse máximo, ele utilizará o disco para armazenar informações que não couberem na memória. Especifique um número maior ou igual a 4.

- Consulte o administrador do sistema para obter o maior valor que pode ser especificado em seu sistema.
- O algoritmo poderá falhar ao localizar o número correto ou especificado de clusters se esse valor for muito baixo.

Padronização de variável. O algoritmo de clusterização funciona com variáveis contínuas padronizadas. Quaisquer variáveis contínuas que não estiverem padronizadas devem ser deixadas como variáveis na lista A Serem Padronizadas. Para economizar tempo e esforço computacional, é possível selecionar quaisquer variáveis contínuas que já tiverem sido padronizadas como variáveis na lista Padronizadas Assumidas.

Opções Avançadas

Critérios de Ajuste de Árvore CF. As seguintes configurações de algoritmo de clusterização se aplicam especificamente à árvore de recursos de cluster (CF) e devem ser alteradas com cuidado:

- **Limite de Mudança da Distância Inicial.** Este é o limite inicial utilizado para crescer a árvore CF. Se inserir um determinado caso em uma folha da árvore CF gerar uma tensão menor que o limite, a folha não será dividida. Se a tensão exceder o limite, a folha será dividida.
- **Máximo de ramificações (por nó folha).** O número máximo de nós filhos que um nó folha pode ter.
- **Profundidade Máxima da Árvore.** O número máximo de níveis que a árvore CF pode ter.
- **Número Máximo de Nós Possíveis.** Isso indica o número máximo de nós de árvore CF que podem potencialmente ser gerados pelo procedimento, com base na função $(b^{d+1} - 1) / (b - 1)$, em que b é o máximo de ramificações e d é a profundidade máxima da árvore. Lembre-se de que sobrepor uma árvore CF grande poderá esgotar os recursos do sistema e prejudicar o desempenho do procedimento. No mínimo, cada nó requer 16 bytes.

Atualização do Modelo de Cluster. Este grupo permite importar e atualizar um modelo de cluster gerado em uma análise anterior. O arquivo de entrada contém a árvore CF em formato XML. Em seguida, o modelo será atualizado com os dados no arquivo ativo. Os nomes de variáveis devem ser selecionados na caixa de diálogo principal na mesma ordem em que eles foram especificados na análise anterior. O arquivo XML permanece inalterado, a menos que você grave especificamente as informações do novo modelo no mesmo nome de arquivo. Consulte o tópico “Saída da Análise de Cluster TwoStep” para obter mais informações

Se uma atualização do modelo de cluster for especificada, as opções referentes à geração da árvore CF que foram especificadas para o modelo original serão utilizadas. Mais especificamente, as configurações de critérios de ajuste de medida de distância, de tratamento de ruído, de alocação de memória ou de árvore CF para o modelo salvo são utilizadas, e todas as configurações para essas opções nas caixas de diálogo são ignoradas.

Nota: Ao executar uma atualização do modelo de cluster, o procedimento supõe que nenhum dos casos selecionados no conjunto de dados ativo foi utilizado para criar o modelo de cluster original. O procedimento também supõe que os casos usados na atualização do modelo são provenientes da mesma população que os casos utilizados para criar o modelo original, ou seja, as médias e variâncias de variáveis contínuas e os níveis de variáveis categóricas são considerados os mesmos nos dois conjuntos de casos. Se os seus conjuntos de casos "novos" e "antigos" forem provenientes de populações heterogêneas, deve-se executar o procedimento de Análise de Cluster TwoStep nos conjuntos de casos combinados para obter os melhores resultados.

Saída da Análise de Cluster TwoStep

Saída. Este grupo fornece opções para exibir os resultados de clusterização.

- **Tabelas dinâmicas.** Os resultados são exibidos em tabelas dinâmicas.
- **Gráficos e tabelas no Visualizador de Modelo.** Os resultados são exibidos no Visualizador de Modelo.
- **Campos de avaliação.** Isso calcula os dados do cluster para as variáveis que não forem utilizadas na criação do cluster. Campos de avaliação podem ser exibidos ao longo dos recursos de entrada no visualizador de modelo ao selecioná-los no subdiálogo Exibir. Campos com valores omissos são ignorados.

Trabalhando com Arquivo de Dados. Este grupo permite salvar as variáveis no conjunto de dados ativo.

- **Criar variável de associação de cluster.** Esta variável contém um número de identificação do cluster para cada caso. O nome dessa variável é *tsc_n*, em que *n* é um número inteiro positivo indicando o ordinal da operação de salvamento do conjunto de dados ativo concluída por este procedimento em uma determinada sessão.

Arquivos XML. O modelo de cluster final e a árvore CF são dois tipos de arquivos de saída que podem ser exportados em formato XML.

- **Exportar modelo final.** O modelo de cluster final é exportado no arquivo especificado em formato XML (PMML). É possível usar esse arquivo de modelo para aplicar as informações de modelo a outros arquivos de dados para propósitos de escoragem.
- **Exportar árvore CF.** Esta opção permite salvar o estado atual da árvore de cluster e atualizá-lo posteriormente utilizando dados mais recentes.

O Visualizador de Cluster

Modelos de cluster são usados normalmente para localizar grupos (ou clusters) de registros semelhantes com base nas variáveis examinadas, onde a semelhança entre membros do mesmo grupo é alta e a semelhança entre membros de grupos diferentes é baixa. Os resultados podem ser usados para identificar associações que não iriam, de outra maneira, ser aparentes. Por exemplo, através da análise de cluster de preferências do cliente, nível de receita e hábitos de compra, pode ser possível identificar os tipos de clientes que são mais propensos a responder a uma campanha de marketing específica.

Existem duas abordagens para interpretar os resultados em uma exibição de cluster:

- Examine clusters para determinar características exclusivas para esse cluster. *Um cluster contém todos os solicitantes de crédito de alta renda? Esse cluster contém mais registros que os outros?*
- Examine campos ao longo de clusters para determinar como os valores são distribuídos entre clusters. *O nível de educação de uma pessoa determina a associação em um cluster? Um escore de risco de crédito alto se distingue entre associação em um cluster ou outro?*

Usando as visualizações principais e as várias visualizações vinculadas no Visualizador de Cluster, é possível ganhar insight para ajudá-lo a responder essas perguntas.

Para ver informações sobre o modelo de cluster, ative (clique duas vezes) no objeto Visualizador de Modelo no Visualizador.

Visualizador de Cluster

O Visualizador de Cluster é composto de dois painéis, a visualização principal à esquerda e a visualização vinculada ou auxiliar à direita. Existem duas visualizações principais:

- Sumarização do Modelo (o padrão). Consulte o tópico “Visualização de Sumarização do Modelo” na página 113 para obter mais informações
- Clusters. Consulte o tópico “Visualização de Clusters” na página 113 para obter mais informações

Existem quatro visualizações vinculadas/auxiliares:

- Importância do Preditor. Consulte o tópico “Visualização de Importância do Preditor de Cluster” na página 115 para obter mais informações
- Tamanhos do Cluster (o padrão). Consulte o tópico “Visualização de Tamanhos de Cluster” na página 115 para obter mais informações
- Distribuição da Célula. Consulte o tópico “Visualização de Distribuição de Célula” na página 115 para obter mais informações
- Comparação do Cluster. Consulte o tópico “Visualização de Comparação do Cluster” na página 115 para obter mais informações

Visualização de Sumarização do Modelo

A visualização de Sumarização do Modelo mostra uma captura instantânea ou uma sumarização do modelo de cluster, incluindo uma medida de Silhueta de coesão e separação do cluster que é sombreada para indicar resultados insatisfatórios, justos ou bons. Essa captura instantânea lhe permite verificar rapidamente se a qualidade é insatisfatória e, nesse caso, você pode decidir retornar para o nó de modelagem para corrigir as configurações de modelo de cluster para produzir um melhor resultado.

Os resultados insatisfatórios, justos e bons são baseados no trabalho de Kaufman e Rousseeuw (1990) relativos à interpretação de estruturas de cluster. Na visualização de Sumarização do Modelo, um bom resultado equivale a dados que refletem a classificação de Kaufman e Rousseeuw como evidência razoável ou forte de estrutura de cluster, resultado justo reflete a sua classificação de evidência fraca e insatisfatório reflete a sua classificação de evidência insignificante.

As médias de medida de silhueta, sobre todos os registros, (B−A) / máx(A,B), em que A é a distância do registro para o seu centro do cluster e B é a distância do registro para o centro do cluster mais próximo ao qual ele não pertence. Um coeficiente de silhueta 1 significaria que todos os casos estão localizados diretamente em seus centros do cluster. Um valor 1 significaria que todos os casos estão localizados nos centros do cluster de algum outro cluster. Um valor 0 significa, em média, que os casos estão equidistantes entre o seu próprio centro do cluster e o outro cluster mais próximo.

A sumarização inclui uma tabela que contém as informações a seguir:

- **Algoritmo.** O algoritmo de clusterização usado, por exemplo, "TwoStep".
- **Variáveis de Entrada.** O número de campos, também conhecido como **entradas** ou **preditores**.
- **Clusters.** O número de clusters na solução.

Visualização de Clusters

A visualização de Clusters contém uma grade de cluster por variáveis que inclui nomes, tamanhos e perfis de cluster para cada cluster.

As colunas na grade contêm as informações a seguir:

- **Cluster.** Os números do cluster criados pelo algoritmo.
- **Rótulo.** Qualquer rótulo aplicado a cada cluster (em branco por padrão). Clique duas vezes na célula para inserir um rótulo que descreve os conteúdos do cluster; por exemplo, "Compradores de carros de luxo".
- **Descrição.** Qualquer descrição dos conteúdos do cluster (em branco por padrão). Clique duas vezes na célula para inserir uma descrição do cluster; por exemplo, "+ de 55 anos de idade, profissionais, que ganham acima de $100.000 por ano".
- **Tamanho.** O tamanho de cada cluster como uma porcentagem da amostra geral de cluster. Cada tamanho de célula dentro da grade exibe uma barra vertical que mostra a porcentagem de tamanho dentro do cluster, uma porcentagem de tamanho em formato numérico e as contagens de caixa do cluster.
- **Variáveis.** As entradas individuais ou preditores, ordenados por importância geral por padrão. Se quaisquer colunas tiverem tamanhos iguais elas serão mostradas em ordenação ascendente dos números do cluster.

  A importância geral da variável é indicada pela cor do sombreamento de segundo plano da célula; a variável mais importante é mais escura; a variável menos importante é sem sombreamento. Um guia acima da tabela indica a importância conectada a cada cor da célula da variável.

Quando você passa o mouse sobre uma célula, o nome/rótulo completo da variável e o valor de importância para a célula é exibido. Informações adicionais podem ser exibidas, dependendo da visualização e do tipo de variável. Na visualização de Centros do Cluster, isso inclui a estatística de célula e o valor da célula; por exemplo: “Média: 4,32”. Para variáveis categóricas a célula mostra o nome da categoria mais frequente (modal) e a sua porcentagem.

Dentro da visualização de Clusters, é possível selecionar várias maneiras para exibir as informações do cluster:

- Transpor clusters e variáveis. Consulte o tópico “Transpor Clusters e Variáveis” para obter mais informações
- Ordenar variáveis. Consulte o tópico “Ordenar Variáveis” para obter mais informações
- Ordenar clusters. Consulte o tópico “Ordenar Clusters” para obter mais informações
- Selecionar conteúdos da célula. Consulte o tópico “Conteúdos da Célula” para obter mais informações

Transpor Clusters e Variáveis: Por padrão, clusters são exibidos como colunas e variáveis são exibidas como linhas. Para inverter essa exibição, clique no botão **Transpor Clusters e Variáveis** à esquerda dos botões **Ordenar Variáveis Por**. Por exemplo, talvez você queira fazer isso quando tiver muitos clusters exibidos, para reduzir a quantia de rolagem horizontal necessária para ver os dados.

Ordenar Variáveis: Os botões **Ordenar Variáveis Por** lhe permitem selecionar como as células da variável são exibidas:

- **Importância Geral.** Esse é o padrão de ordenação. Variáveis são ordenadas em ordem decrescente de importância geral e a ordenação é a mesma ao longo de clusters. Se qualquer variável tiver valores de importância ligados, as variáveis ligadas são listadas em ordenação ascendente dos nomes de variáveis.
- **Importância Dentro do Cluster.** Variáveis são ordenadas em relação à sua importância para cada cluster. Se qualquer variável tiver valores de importância ligados, as variáveis ligadas são listadas em ordenação ascendente dos nomes de variáveis. Quando essa opção é escolhida, a ordenação geralmente varia ao longo de clusters.
- **Nome.** Variáveis são ordenadas por nome em ordem alfabética.
- **Ordem de dados.** Variáveis são ordenadas por sua ordem no conjunto de dados.

Ordenar Clusters: Por padrão, clusters são ordenados em ordem decrescente de tamanho. Os botões **Ordenar Clusters Por** permitem que você os ordene por nome em ordem alfabética ou, se você criou rótulos exclusivos, em ordem alfanumérica de rótulo em vez disso.

Variáveis que têm o mesmo rótulo são ordenadas por nome do cluster. Se clusters forem ordenados por rótulo e você editar o rótulo de um cluster, a ordenação será atualizada automaticamente.

Conteúdos da Célula: Os botões de **Células** lhe permitem mudar a exibição dos conteúdos da célula para variáveis e campos de avaliação.

- **Centros do Cluster.** Por padrão, células exibem nomes/rótulos de variável e a tendência central para cada combinação de cluster/variável. A média é mostrada para campos contínuos e o modo (categoria que ocorre mais frequentemente) com porcentagem de categoria para campos categóricos.
- **Distribuições Absolutas.** Mostra nomes/rótulos de variável e distribuições absolutas das variáveis dentro de cada cluster. Para variáveis categóricas, a exibição mostra gráficos de barras sobrepostos com categorias ordenadas em ordem crescente dos valores dos dados. Para variáveis contínuas, a exibição mostra um gráfico de densidade suave que usa os mesmos terminais e intervalos para cada cluster.

  A exibição de cor vermelha sólida mostra a distribuição de cluster, enquanto a exibição mais pálida representa os dados gerais.
- **Distribuições Relativas.** Mostra nomes/rótulos de variável e distribuições relativas nas células. Em geral, as exibições são semelhantes àquelas mostradas para distribuições absolutas, exceto que distribuições relativas são exibidas em vez disso.

  A exibição de cor vermelha sólida mostra a distribuição de cluster, enquanto a exibição mais pálida representa os dados gerais.
- **Visualização Básica.** Onde houver muitos clusters, pode ser difícil ver todos os detalhes sem rolagem. Para reduzir a quantia de rolagem, selecione essa visualização para mudar a exibição para uma versão mais compacta da tabela.

Visualização de Importância do Preditor de Cluster

A visualização de Importância do Preditor mostra a importância relativa de cada campo na estimativa do modelo.

Visualização de Tamanhos de Cluster

A visualização de Tamanhos de Cluster mostra um gráfico de pizza que contém cada cluster. O tamanho da porcentagem de cada cluster é mostrado em cada fatia; passe o mouse sobre cada fatia para exibir a contagem dessa fatia.

Abaixo do gráfico, uma tabela lista as informações de tamanho a seguir:

- O tamanho do menor cluster (ambas: uma contagem e uma porcentagem do todo).
- O tamanho do maior cluster (ambas: uma contagem e uma porcentagem do todo).
- A razão de tamanho do maior cluster para o menor cluster.

Visualização de Distribuição de Célula

A visualização de Distribuição de Célula mostra um gráfico expandido, mais detalhado, da distribuição dos dados para qualquer célula de variável que você selecionar na tabela no painel principal de Clusters.

Visualização de Comparação do Cluster

A visualização de Comparação do Cluster consiste em um layout de estilo em grade, com variáveis nas linhas e clusters selecionados nas colunas. Essa visualização o ajuda a entender melhor os fatores que compõem os clusters; ela também lhe permite ver diferenças entre clusters não apenas conforme comparados com os dados gerais, mas uns com os outros.

Para selecionar clusters para exibição, clique na parte superior da coluna de cluster no painel principal de Clusters. Use Ctrl-clique ou pressione Shift e clique para selecionar ou cancelar a seleção de mais de um cluster para comparação.

Nota: É possível selecionar até cinco clusters para exibição.

Clusters são mostrados na ordem em que eles foram selecionados, enquanto a ordem de campos é determinada pela opção **Ordenar Variáveis Por**. Quando você seleciona **Importância Dentro do Cluster**, os campos são sempre ordenados por importância geral.

Os gráficos de segundo plano mostram as distribuições gerais de cada variável:

- Variáveis categóricas são mostradas como gráficos de pontos, em que o tamanho do ponto indica a categoria mais frequente/modal para cada cluster (por variável).
- Variáveis contínuas são exibidas como diagramas de caixa, que mostram medianas gerais e as amplitudes interquartis.

Sobrepostos nessas visualizações de segundo plano estão diagramas de caixa para clusters selecionados:

- Para variáveis contínuas, marcadores de ponto quadrado e linhas horizontais indicam a amplitude mediana e interquartil para cada cluster.
- Cada cluster é representado por uma cor diferente, mostrada na parte superior da visualização.

Navegando no Visualizador de Cluster

O Visualizador de Cluster é uma exibição interativa. É possível:

- Selecionar um campo ou cluster para visualizar mais detalhes.
- Comparar clusters para selecionar itens de interesse.
- Alterar a exibição.
- Transpor eixos.

Usando as Barras de Ferramentas

Você controla as informações mostradas em ambos os painéis: esquerdo e direito usando as opções da barra de ferramentas. É possível mudar a orientação da exibição (de cima para baixo, da esquerda para a direita ou da direita para a esquerda) usando os controles da barra de ferramentas. Além disso, também é possível reconfigurar o visualizador para as configurações padrão e abrir uma caixa de diálogo para especificar os conteúdos da visualização de Clusters no painel principal.

As opções **Ordenar Variáveis Por**, **Ordenar Clusters Por**, **Células** e **Exibir** estão disponíveis somente quando você seleciona a visualização de **Clusters** no painel principal. Consulte o tópico "Visualização de Clusters" na página 113 para obter mais informações

Tabela 2. Ícones da barra de ferramentas.

Ícone	Tópico
	Consulte Transpor Clusters e Variáveis
	Consulte Ordenar Variáveis Por
	Consulte Ordenar Clusters Por
	Consulte Células

Exibição de Visualização de Cluster de Controle

Para controlar o que é mostrado na visualização de Clusters no painel principal, clique no botão **Exibir**; o diálogo Exibir é aberto.

Variáveis. Selecionado por padrão. Para ocultar todas as variáveis de entrada, cancele a seleção da caixa de seleção.

Campos de Avaliação. Escolha os campos de avaliação (campos não usados para criar o modelo de cluster, mas enviados para o visualizador de modelo para avaliar os clusters) para exibir; nenhum é mostrado por padrão. *Nota* O campo de avaliação deve ser uma sequência de caracteres com mais de um valor. Essa caixa de seleção estará indisponível se nenhum campo de avaliação estiver disponível.

Descrições do Cluster. Selecionado por padrão. Para ocultar todas as células de descrição do cluster, cancele a seleção da caixa de seleção.

Tamanhos do Cluster. Selecionado por padrão. Para ocultar todas as células de tamanho do cluster, cancele a seleção da caixa de seleção.

Número Máximo de Categorias. Especifica o número máximo de categorias para exibir em gráficos de variáveis categóricas; o padrão é 20.

Filtrando Registros

Se você quiser saber mais sobre os casos em um determinado cluster ou grupo de clusters, é possível selecionar um subconjunto de registros para posterior análise com base nos clusters selecionados.

1. Selecione os clusters na visualização de Cluster do Visualizador de Cluster. Para selecionar vários clusters, use Ctrl-clique
2. Nos menus, escolha:

   Gerar > **Filtrar Registros...**

3. Insira um nome de variável de filtro. Registros a partir dos clusters selecionados receberão um valor 1 para esse campo. Todos os outros registros receberão um valor 0 e serão excluídos de análises subsequentes até você mudar o status do filtro.
4. Clique em **OK**.

Capítulo 25. Análise de clusters hierárquica

Este procedimento tenta identificar grupos de casos (ou variáveis) relativamente homogêneos com base nas características selecionadas, utilizando um algoritmo que inicia com cada caso (ou variável) em um cluster separado e combina clusters até restar somente um. É possível analisar variáveis brutas, ou escolher a partir de uma variedade de transformações de padronização. Medidas de distância ou de similaridade são geradas pelo procedimento Proximidades. As estatísticas são exibidas em cada estágio para ajudá-lo a selecionar a melhor solução.

Exemplo. Há grupos identificáveis de programas de televisão que atraem públicos semelhantes para dentro de cada grupo? Com a análise de cluster hierárquica, é possível agrupar programas de televisão (casos) em grupos homogêneos com base nas características do telespectador. Esse processo pode ser utilizado para identificar segmentos para marketing. Ou é possível agrupar cidades (casos) em grupos homogêneos, de modo que cidades comparáveis possam ser selecionadas para testar as várias estratégias de marketing.

Estatísticas. Planejamento de aglomeração, matriz de distância (ou similaridade) e associação de cluster para uma solução única ou para um intervalo de soluções. Gráficos: dendrogramas e gráficos icicle.

Considerações de Dados de Análise de Cluster Hierárquica

Dados. As variáveis podem ser quantitativas, binárias ou dados de contagem. O ajuste de escala de variáveis é uma consideração importante - as diferenças no ajuste de escala podem afetar uma ou mais de suas soluções de cluster. Se as suas variáveis possuírem diferenças grandes no ajuste de escala (por exemplo, uma variável é medida em dólares e a outra é medida em anos), considere padronizá-las (esse processo pode ser feito automaticamente pelo procedimento Ajuste de Escala Hierárquica).

Ordem de caso. Se distâncias ou similaridades empatadas existirem nos dados de entrada ou ocorrerem entre clusters atualizados durante a junção, a solução de cluster resultante poderá depender da ordem dos casos no arquivo. Talvez você queira obter várias soluções diferentes com casos classificados em diferentes ordens aleatórias para verificar a estabilidade de uma determinada solução.

Suposições. As medidas de distância ou de similaridade utilizadas devem ser apropriadas para os dados analisados (consulte o procedimento Proximidades para obter mais informações sobre as opções de medidas de distância e de similaridade). Além disso, deve-se incluir todas as variáveis relevantes em sua análise. Omissão de variáveis influentes pode resultar em uma solução enganosa. Como a análise de cluster hierárquica é um método exploratório, os resultados devem ser tratados como tentativa até que eles sejam confirmados com uma amostra independente.

Para Obter uma Análise de Cluster Hierárquica

1. Nos menus, escolha:

   Analisar > **Classificar** > **Cluster Hierárquico...**
2. Se estiver armazenando casos em cluster, selecione pelo menos uma variável numérica. Se estiver armazenando variáveis em cluster, selecione pelo menos três variáveis numéricas.

Opcionalmente, é possível selecionar uma variável de identificação para rotular casos.

Método de Análise de Cluster Hierárquica

Método de Cluster. As alternativas disponíveis são ligação entre grupos, ligação dentro de grupos, vizinho mais próximo, vizinho mais distante, clusterização de centroide, clusterização de mediana e o método de Ward.

Medida. Permite especificar a medida de distância ou de similaridade ser utilizada em cluster. Selecione o tipo de dados e a medida de distância ou de similaridade adequada:

- **Intervalo.** As alternativas disponíveis são distância euclidiana, distância euclidiana quadrática, cosseno, correlação de Pearson, Chebychev, bloco, Minkowski e customizado.
- **Contagens.** As alternativas disponíveis são medida de qui-quadrado e medida de fi-quadrado.
- **Binário.** As alternativas disponíveis são distância euclidiana, distância euclidiana quadrática, diferença de tamanho, diferença padrão, variância, dispersão, forma, correspondência simples, correlação fi de 4 pontos, lambda, *D* de Anderberg, divisão, 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* de Yule e *Q* de Yule.

Transformar valores. Permite padronizar valores de dados para casos ou valores antes de calcular proximidades (não disponível para dados binários). Os métodos de padronização disponíveis são escores *z*, intervalo -1 a 1, intervalo 0 a 1, magnitude máxima de 1, média de 1 e desvio padrão de 1.

Transformar Medidas. Permite transformar os valores gerados pela medida de distância. Eles são aplicados após a medida de distância ter sido calculada. As alternativas disponíveis são valores absolutos, sinal de mudança e reajuste de escala para o intervalo 0-1.

Estatísticas de Análise de Cluster Hierárquica

Planejamento de aglomeração. Exibe os casos ou clusters combinados em cada estágio, as distâncias entre os casos ou clusters que estão sendo combinados e o último nível do cluster no qual um caso (ou variável) se uniu ao cluster.

Matriz de proximidade. Fornece as distâncias ou similaridades entre os itens.

Associação de Cluster. Exibe o cluster ao qual cada caso é designado em um ou mais estágios na combinação de clusters. As opções disponíveis são solução única e intervalo de soluções.

Gráficos de Análise de Cluster Hierárquica

Dendrograma. Exibe um *dendrograma*. Dendrogramas podem ser utilizados para avaliar a coesão dos clusters formados e pode fornecer informações sobre o número apropriado de clusters a serem mantidos.

Icicle. Exibe um *gráfico icicle*, incluindo todos os clusters ou um intervalo de clusters especificado. Os gráficos icicle exibem informações sobre como os casos são combinados em clusters em cada iteração da análise. A orientação permite selecionar um gráfico vertical ou horizontal.

Análise de Cluster Hierárquica para Salvar Novas Variáveis

Associação de Cluster. Permite salvar as associações de cluster para uma solução única ou um intervalo de soluções. As variáveis salvas podem, então, ser utilizadas em análises subsequentes para explorar outras diferenças entre os grupos.

Recursos Adicionais da Sintaxe do Comando CLUSTER

O procedimento Cluster Hierárquico utiliza a sintaxe de comando do `CLUSTER`. O idioma da sintaxe de comando também permite:

- Utilizar vários métodos de clusterização em uma análise única.
- Ler e analisar uma matriz de proximidade.
- Gravar uma matriz de proximidade no disco para análise posterior.
- Especificar quaisquer valores para potência e raiz na medida de distância customizada (Potência).
- Especificar nomes para variáveis salvas.

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 26. Análise de cluster por K-médias

Este procedimento tenta identificar grupos de casos relativamente homogêneos com base em características selecionadas, utilizando um algoritmo que possa manipular números grandes de casos. No entanto, o algoritmo requer que você especifique o número de clusters. É possível especificar centros do cluster iniciais se você souber essas informações. É possível selecionar um dos dois métodos para classificar casos, atualizar os centros do cluster interativamente ou somente classificar. É possível salvar a associação de cluster, informações de distância e centros do cluster finais. Opcionalmente, é possível especificar uma variável cujos valores são utilizados para rotular a saída entre casos. Também é possível solicitar estatísticas *F* de análise de variância. Embora essas estatísticas sejam oportunistas (o procedimento tenta formar grupos que diferem), o tamanho relativo das estatísticas fornece informações sobre a contribuição de cada variável para a separação dos grupos.

Exemplo. Quais são alguns grupos identificáveis de programas de televisão que atraem públicos semelhantes para dentro de cada grupo? Com a análise de cluster por *k*-médias, é possível agrupar programas de televisão (casos) em *k* grupos homogêneos com base nas características do telespectador. Esse processo pode ser utilizado para identificar segmentos para marketing. Ou é possível agrupar cidades (casos) em grupos homogêneos, de modo que cidades comparáveis possam ser selecionadas para testar as várias estratégias de marketing.

Estatísticas. Solução completa: centros do cluster iniciais, tabela ANOVA. Cada caso: informações do cluster, distância do centro do cluster.

Considerações de Dados de Análise de Cluster por K-Médias

Dados. As variáveis devem ser quantitativas no nível de intervalo ou de razão. Se suas variáveis forem binárias ou contagens, use o procedimento de Análise de Cluster Hierárquica.

Caso e ordem inicial do centro do cluster. O algoritmo padrão para escolher os centros do cluster iniciais não é invariável para ordenação de caso. A opção **Utilizar médias de execução** na caixa de diálogo Iterar torna a solução resultante potencialmente dependente da ordem do caso, não importa como os centros do cluster iniciais forem escolhidos. Se estiver utilizando qualquer um destes métodos, talvez você queira obter várias soluções diferentes com casos classificados em diferentes ordens aleatórias para verificar a estabilidade de uma determinada solução. Especificar centros do cluster iniciais e não utilizar a opção **Utilizar médias de execução** evitam problemas relacionados à ordem do caso. No entanto, a ordenação dos centros do cluster iniciais poderá afetar a solução se houver distâncias empatadas entre os casos e os centros do cluster. Para avaliar a estabilidade de uma determinada solução, será possível comparar os resultados das análises com diferentes permutações dos valores do centro iniciais.

Suposições. As distâncias são calculadas usando a distância euclidiana simples. Se desejar utilizar outra distância ou medida de similaridade, use o procedimento de Análise de Cluster Hierárquica. O ajuste de escala de variáveis é uma consideração importante. Se as suas variáveis forem medidas em escalas diferentes (por exemplo, uma variável é expressa em dólares e outra variável é expressa em anos), os resultados poderão ser enganosos. Nesses casos, deve-se considerar a padronização de suas variáveis antes de executar a análise de cluster por *k*-médias (esta tarefa pode ser feita no procedimento Descritivos). O procedimento supõe que você selecionou o número apropriado de clusters e que todas as variáveis relevantes foram incluídas. Se você escolheu um número inadequado de clusters ou omitiu variáveis importantes, os resultados poderão ser enganosos.

Para Obter uma Análise de Cluster por K-Médias

1. Nos menus, escolha:

   Analisar > **Classificar** > **Cluster de K-Médias...**

2. Selecione as variáveis a serem utilizadas na análise de cluster.
3. Especifique o número de clusters. (O número de clusters deve ser pelo menos 2 e não deve ser maior que o número de casos no arquivo de dados).
4. Selecione **Iterar e classificar** ou **Somente classificar**.
5. Opcionalmente, selecione uma variável de identificação para rotular casos.

Eficiência de Análise de Cluster por K-Médias

O comando de análise de cluster por *k*-médias é eficaz principalmente porque ele não calcula as distâncias entre todos os pares de casos, assim como fazem muitos algoritmos de cluster, incluindo o algoritmo que é utilizado pelo comando de clusterização hierárquico.

Para obter o máximo de eficiência, pegue uma amostra de casos e selecione o método **Iterar e classificar** para determinar os centros do cluster. Selecione **Gravar final como**. Em seguida, restaure o arquivo de dados inteiro e selecione **Somente classificar** como o método e selecione **Leitura inicial de** para classificar o arquivo inteiro usando os centros que são estimados a partir da amostra. É possível gravar e ler a partir de um arquivo ou de um conjunto de dados. Conjuntos de dados estão disponíveis para uso subsequente na mesma sessão, mas não são salvos como arquivos, a menos que sejam salvos explicitamente antes do término da sessão. Os nomes do conjunto de dados devem estar de acordo com as regras de nomenclatura de variáveis. Veja o tópico para obter mais informações.

Iteração de Análise de Cluster por K-Médias

Nota: Estas opções estarão disponíveis apenas se selecionar o método **Iterar e classificar** a partir da caixa de diálogo Análise de Cluster por K-Médias.

Máximo de Iterações. Limita o número de iterações no algoritmo *k*-médias. A iteração parará após este número de iterações, mesmo se o critério de convergência não estiver satisfeito. Este número deve ser entre 1 e 999.

Para reproduzir o algoritmo utilizado pelo comando Cluster Rápido anterior à versão 5.0, configure **Máximo de Iterações** para 1.

Critério de Convergência. Determina quando iteração é interrompida. Ele representa uma proporção da distância mínima entre os centros do cluster iniciais, de modo que ele deve ser maior que 0, mas não maior que 1. Se o critério for igual a 0,02, por exemplo, a iteração parará quando uma iteração completa não mover nenhum dos centros do cluster por uma distância de mais de 2% da menor distância entre qualquer um dos centros do cluster iniciais.

Utilizar médias de execução. Permite solicitar que os centros do cluster sejam atualizados após cada caso ser designado. Se você não selecionar essa opção, novos centros do cluster serão calculados após todos os casos terem sido designados.

Salvar Análise de Cluster por K-Médias

É possível salvar informações sobre a solução como novas variáveis a serem usadas em análises subsequentes:

Associação de cluster. Cria uma nova variável indicando a associação de cluster final de cada caso. Os valores da nova variável variam de 1 até o número de clusters.

Distância do centro do cluster. Cria uma nova variável indicando a distância euclidiana entre cada caso e seu centro de classificação.

Opções de Análise de Cluster por K-médias

Estatísticas. É possível selecionar as seguintes estatísticas: centros do cluster iniciais, tabela de ANOVA e informações do cluster para cada caso.

- *Centros do cluster iniciais* Primeira estimativa das médias da variável para cada um dos clusters. Por padrão, um número de casos bem espaçados igual ao número de clusters é selecionado a partir dos dados. Os centros do cluster iniciais são utilizados para uma primeira rodada de classificação e, em seguida, são atualizados.
- *Tabela de ANOVA*. Exibe uma tabela de análise de variância que inclui testes F univariados para cada variável de clusterização. Os testes F são apenas descritivos e as probabilidades resultantes não devem ser interpretadas. A tabela ANOVA não será exibida se todos os casos forem designados para um único cluster.
- *Informações de cluster para cada caso*. Exibe para cada caso a designação do cluster final e a distância euclidiana entre o caso e o centro do cluster utilizadas para classificar o caso. Exibe também a distância euclidiana entre os centros do cluster final.

Valores omissos. As opções disponíveis são **Excluir listwise dos casos** ou **Excluir Casos Entre Pares**.

- **Excluir listwise dos casos.** Exclui da análise casos com valores omissos para qualquer variável de clusterização.
- **Excluir casos entre pares.** Designa casos para clusters com base nas distâncias que são calculadas a partir de todas as variáveis com valores não omissos.

Recursos Adicionais do Comando QUICK CLUSTER

O procedimento Cluster de K-Médias utiliza a sintaxe de comando `QUICK CLUSTER`. O idioma da sintaxe de comando também permite:

- Aceitar os primeiros *k* casos como os centros do cluster iniciais, evitando, assim, a passagem de dados que é normalmente utilizada para estimá-los.
- Especifique os centros do cluster iniciais diretamente como parte da sintaxe de comando.
- Especificar nomes para variáveis salvas.

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 27. Testes Não Paramétricos

Testes não paramétricos fazem premissas mínimas sobre a distribuição subjacente dos dados. Os testes que estão disponíveis nesses diálogos podem ser agrupados em três amplas categorias com base em como os dados são organizados:

- Um teste de uma amostra analisa um campo.
- Um teste para amostras relacionadas compara dois ou mais campos para o mesmo conjunto de casos.
- Um teste de amostras independentes analisa um campo que é agrupado por categorias de um outro campo.

Testes Não paramétricos de uma amostra

Testes não paramétricos de uma amostra identificam diferenças em campos únicos usando um ou mais testes não paramétricos. Os testes não paramétricos não presumem que seus dados sigam a distribuição normal.

Qual é o seu objetivo? Os objetivos lhe permitem especificar rapidamente configurações de teste diferentes, mas usadas frequentemente.

- **Comparar automaticamente dados observados com hipotéticos.** Esse objetivo aplica o teste Binomial em campos categóricos com apenas duas categorias, o teste Qui-Quadrado em todos os outros campos categóricos e o teste de Kolmogorov-Smirnov em campos contínuos.
- **Sequência de teste para aleatoriedade.** Esse objetivo usa o teste de Execuções para testar a sequência observada de valores dos dados para aleatoriedade.
- **Análise customizada.** Quando você desejar corrigir manualmente as configurações de teste na guia Configurações, selecione essa opção. Observe que essa configuração será selecionada automaticamente se você subsequentemente fizer mudanças em opções na guia Configurações que forem incompatíveis com o objetivo selecionado atualmente.

Para Obter Testes Não Paramétricos de Uma Amostra

Nos menus, escolha:

Analisar > **Testes Não Paramétricos** > **Uma Amostra...**

1. Clique em **Executar**.

Como opção, você pode:

- Especificar um objetivo na guia Objetivo.
- Especificar designações de campo na guia Campos.
- Especificar configurações de especialista na guia Configurações.

Guia Campos

A guia Campos especifica quais campos devem ser testados.

Usar funções predefinidas. Essa opção usa informação de campo. Todos os campos com uma função predefinida como Entrada, Resposta ou Ambos serão usados como campos de teste. Pelo menos um campo de teste é necessário.

Usar designações de campo customizado. Essa opção lhe permite substituir papéis do campo. Após selecionar essa opção, especifique os campos abaixo:

- **Campos de Teste.** Selecione um ou mais campos.

Guia Configurações

A guia Configurações abrange vários grupos diferentes de configurações que é possível modificar para ajustar com precisão como o algoritmo processa os seus dados. Se você fizer qualquer mudança nas configurações padrão que for incompatível com os objetivos selecionados atualmente, a guia Objetivo será atualizada automaticamente para selecionar a opção **Customizar análise**.

Escolher Testes

Essas configurações especificam os testes a serem executados nos campos especificados na guia Campos.

Escolher automaticamente os testes com base nos dados. Essa configuração aplica o teste Binomial em campos categóricos com apenas duas categorias válidas (não omissas), o teste Qui-Quadrado em todos os outros campos categóricos e o teste de Kolmogorov-Smirnov em campos contínuos.

Customizar testes. Essa configuração permite que você escolha testes específicos a serem executados.

- **Comparar probabilidade binária observada com hipotética (Teste binomial).** O Teste binomial pode ser aplicado a todos os campos. Isso produz um teste de uma amostra se a distribuição observada de um campo de sinalização (um campo categórico com apenas duas categorias) é a mesma que o que é esperado a partir de uma distribuição binomial especificada. Além disso, é possível solicitar intervalos de confiança. Consulte "Opções de Testes Binomiais" para obter detalhes sobre as configurações de teste.
- **Comparar probabilidades observadas com hipotéticas (Teste Qui-Quadrado).** O teste Qui-Quadrado é aplicado a campos nominais e ordinais. Isso produz um teste de uma amostra que calcula uma estatística qui-quadrado baseada nas diferenças entre frequências observadas e esperadas de categorias de um campo. Consulte "Opções de Testes Qui-Quadrado" na página 129 para obter detalhes sobre as configurações de teste.
- **Testar distribuição observada com relação à hipotética (Teste de Kolmogorov-Smirnov).** O teste de Kolmogorov-Smirnov é aplicado a campos contínuos e ordinais. Isso produz um teste de uma amostra de se a função de distribuição acumulativa de amostra para um campo é homogênea com uma distribuição uniforme, normal, de Poisson ou exponencial. Consulte "Opções de Kolmogorov-Smirnov" na página 129 para obter detalhes sobre as configurações de teste.
- **Comparar mediana com hipotética (Teste dos postos sinalizados de Wilcoxon).** O teste dos postos sinalizados de Wilcoxon é aplicado a campos contínuos e ordinais. Isso produz um teste de uma amostra de valor médio de um campo. Especifique um número como a mediana hipotética.
- **Sequência de teste para aleatoriedade (Teste de execuções).** O Teste de execuções é aplicado a todos os campos. Isso produz um teste de uma amostra de se a sequência de valores de um campo dicotomizado é aleatória. Consulte "Opções de Teste de Execuções" na página 129 para obter detalhes sobre as configurações de teste.

Opções de Testes Binomiais: O teste binomial destina-se a campos de sinalização (campos categóricos com apenas duas categorias), mas é aplicado a todos os campos usando regras para definir "sucesso".

Proporção hipotética. Isso especifica a proporção esperada de registros definidos como "sucessos" ou *p*. Especifique um valor maior que 0 e menor que 1. O padrão é 0,05.

Intervalo de Confiança. Os métodos a seguir para calcular intervalos de confiança para dados binários estão disponíveis:

- **Clopper-Pearson (exato).** Um intervalo exato baseado na distribuição binomial acumulativa.
- **Jeffreys.** Um intervalo Bayesiano baseado na distribuição posterior de *p* usando as informações a priori de Jeffreys.
- **Razão de verossimilhança.** Um intervalo baseado na função de probabilidade para *p*.

Definir Sucesso para Campos Categóricos. Isso especifica como "sucesso", o(s) valor(es) de dados testado(s) com relação à proporção hipotética, é definido para campos categóricos.

- **Usar a primeira categoria localizada em dados** executa o teste binomial usando o primeiro valor localizado na amostra para definir "sucesso". Essa opção é aplicável somente em campos nominais ou ordinais com apenas dois valores; todos os outros campos categóricos especificados na guia Campos em que essa opção é usada não serão testados. Este é o padrão.
- **Especificar valores de sucesso** executa o teste binomial usando a lista especificada de valores para definir "sucesso". Especifique uma lista de sequências de caracteres ou de valores numéricos. Os valores na lista não precisam estar presentes na amostra.

Definir Sucesso para Campos Contínuos. Isso especifica como "sucesso", o(s) valor(es) de dados testado(s) com relação ao valor de teste, é definido para campos contínuos. Sucesso é definido como valores iguais ou inferiores a um ponto de corte.

- **Ponto médio da amostra** configura o ponto de corte na média dos valores mínimo e máximo.
- **Ponto de corte customizado** permite que você especifique um valor para o ponto de corte.

Opções de Testes Qui-Quadrado: Todas as categorias têm probabilidade igual. Isso produz frequências iguais entre todas as categorias na amostra. Esse é o padrão.

Customizar probabilidade esperada. Isso lhe permite especificar frequências desiguais para uma lista de categorias especificada. Especifique uma lista de sequências de caracteres ou de valores numéricos. Os valores na lista não precisam estar presentes na amostra. Na coluna **Categoria**, especifique valores de categoria. Na coluna **Frequência Parente**, especifique um valor maior que 0 para cada categoria. Frequências customizadas são tratadas como índices de modo que, por exemplo, especificar frequências 1, 2 e 3 é equivalente a especificar frequências 10, 20 e 30 e ambas especificam que é esperado que 1/6 dos registros caiam na primeira categoria, 1/3 na segunda e 1/2 na terceira. Quando probabilidades customizadas esperadas são especificadas, os valores de categoria customizados devem incluir todos os valores do campo nos dados; caso contrário, o teste não será executado para esse campo.

Opções de Kolmogorov-Smirnov: Esse diálogo especifica quais distribuições devem ser testadas e os parâmetros das distribuições hipotéticas.

Normal. Usar dados de amostra usa a média observada e o desvio padrão, **Customizado** permite que você especifique valores.

Uniforme. Usar dados de amostra usa o mínimo e máximo observados, **Customizado** permite que você especifique valores.

Exponencial. Média da amostra usa a média observada, **Customizado** permite que você especifique valores.

Poisson. Média da amostra usa a média observada, **Customizado** permite que você especifique valores.

Opções de Teste de Execuções: O teste de execuções destina-se a campos de sinalização (campos categóricos com apenas duas categorias), mas pode ser aplicado a todos os campos usando regras para definir os grupos.

Definir Grupos para Campos Categóricos. As opções a seguir estão disponíveis:

- **Existem somente 2 categorias na amostra** executa o teste de execuções usando os valores localizados na amostra para definir os grupos. Essa opção é aplicável somente em campos nominais ou ordinais com apenas dois valores; todos os outros campos categóricos especificados na guia Campos em que essa opção é usada não serão testados.
- **Recodificar dados em 2 categorias** executa o teste de execuções usando a lista especificada de valores para definir um dos grupos. Todos os outros valores na amostra definem o outro grupo. Os valores na lista não precisam estar todos presentes na amostra, mas pelo menos um registro deve estar em cada grupo.

Definir Ponto de Corte para Campos Contínuos. Isso especifica como os grupos são definidos para campos contínuos. O primeiro grupo é definido como valores iguais ou inferiores a um ponto de corte.

- **Mediana de amostra** configura o ponto de corte na mediana de amostra.
- **Média de amostra** configura o ponto de corte na média de amostra.
- **Customizado** permite que você especifique um valor para o ponto de corte.

Opções de Teste

Nível de Significância. Isso especifica o nível de significância (alpha) para todos os testes. Especifique um valor numérico entre 0 e 1. 0,05 é o padrão.

Intervalo de confiança (%). Isso especifica o nível de confiança para todos os intervalos de confiança produzidos. Especifique um valor numérico entre 0 e 100. 95 é o padrão.

Casos Excluídos. Isso especifica como determinar a base de caso para testes.

- **Excluir casos listwise** significa que registros com valores omissos para qualquer campo que é nomeado na guia Campos são excluídos de todas as análises.
- **Excluir teste de casos por teste** significa que registros com valores omissos para um campo que é usado para um teste específico são omitidos desse teste. Quando diversos testes são especificados na análise, cada teste é avaliado separadamente.

Valores Omissos de Usuário

Valores Omissos de Usuário para Campos Categóricos. Campos categóricos devem ter valores válidos para um registro a ser incluído na análise. Esses controles permitem que você decida se os valores omissos de usuário são tratados como válidos entre os campos categóricos. Valores omissos do sistema e valores omissos para campos contínuos são sempre tratados como inválidos.

Variáveis Adicionais de Comando de NPTESTS

O idioma da sintaxe de comando também permite:

- Especificar testes de uma amostra, de amostras independentes e de amostras relacionadas em uma única execução do procedimento.

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Testes Não Paramétricos de Amostras Independentes

Testes não paramétricos de amostras independentes identificam diferenças entre dois ou mais grupos usando um ou mais testes não paramétricos. Os testes não paramétricos não presumem que seus dados sigam a distribuição normal.

Qual é o seu objetivo? Os objetivos lhe permitem especificar rapidamente configurações de teste diferentes, mas usadas frequentemente.

- **Comparar automaticamente distribuições ao longo de grupos.** Esse objetivo aplica-se ao teste U de Mann-Whitney para dados com 2 grupos ou ao ANOVA de 1 via de Kruskal-Wallis para dados com grupos *k* .
- **Comparar medianas ao longo de grupos.** Esse objetivo usa o Teste de mediana para comparar as medianas observadas ao longo de grupos.
- **Análise customizada.** Quando você desejar corrigir manualmente as configurações de teste na guia Configurações, selecione essa opção. Observe que essa configuração será selecionada automaticamente se você subsequentemente fizer mudanças em opções na guia Configurações que forem incompatíveis com o objetivo selecionado atualmente.

Para Obter Testes Não Paramétricos de Amostras Independentes

Nos menus, escolha:

Analisar > **Testes Não Paramétricos** > **Amostras Independentes...**

1. Clique em **Executar**.

Como opção, você pode:

- Especificar um objetivo na guia Objetivo.
- Especificar designações de campo na guia Campos.
- Especificar configurações de especialista na guia Configurações.

Guia Campos

A guia Campos especifica quais campos devem ser testados e o campo usado para definir grupos.

Usar funções predefinidas. Essa opção usa informação de campo. Todos os campos contínuos e ordinais com uma função predefinida como Resposta ou Ambos serão usados como campos de teste. Se houver um campo categórico único com uma função predefinida como Entrada, ele será usado como um campo de agrupamento. Caso contrário, nenhum campo de agrupamento será usado por padrão e deve-se usar designações de campo customizado. Pelo menos um campo de teste e um campo de agrupamento são necessários.

Usar designações de campo customizado. Essa opção lhe permite substituir papéis do campo. Após selecionar essa opção, especifique os campos abaixo:

- **Campos de Teste.** Selecione um ou mais campos contínuos ou ordinais.
- **Grupos.** Selecione um campo categórico.

Guia Configurações

A guia Configurações abrange vários grupos diferentes de configurações que é possível modificar para ajustar com precisão como o algoritmo processa os seus dados. Se você fizer qualquer mudança nas configurações padrão que for incompatível com os objetivos selecionados atualmente, a guia Objetivo será atualizada automaticamente para selecionar a opção **Customizar análise**.

Escolher Testes

Essas configurações especificam os testes a serem executados nos campos especificados na guia Campos.

Escolher automaticamente os testes com base nos dados. Essa configuração aplica-se ao teste U de Mann-Whitney para dados com 2 grupos ou ao ANOVA de 1 via de Kruskal-Wallis para dados com grupos *k* .

Customizar testes. Essa configuração permite que você escolha testes específicos a serem executados.

- **Comparar Distribuições ao longo de Grupos.** Isso produz testes de amostras independentes de se as amostras são da mesma população.

  Mann-Whitney U (2 amostras) usa o ranqueamento de cada caso para testar se os grupos são desenhados a partir da mesma população. O primeiro valor em ordem crescente do campo de agrupamento define o primeiro grupo e o segundo define o segundo grupo. Se o campo de agrupamento tiver mais de dois valores, esse teste não será produzido.

  Kolmogorov-Smirnov (2 amostras) é sensível a qualquer diferença em mediana, dispersão, assimetria e assim por diante, entre as duas distribuições. Se o campo de agrupamento tiver mais de dois valores, esse teste não será produzido.

  Sequência de teste para aleatoriedade (Wald-Wolfowitz para 2 amostras) produz um teste de execuções com associação ao grupo como o critério. Se o campo de agrupamento tiver mais de dois valores, esse teste não será produzido.

ANOVA de 1 via de Kruskal-Wallis (amostras k) é uma extensão do teste U de Mann-Whitney e o analógico não paramétrico de análise de variância para um fator. É possível opcionalmente solicitar várias comparações das amostras *k*, todas as várias comparações **entre pares** ou comparações **de redução de stepwise**.

Teste para alternativas ordenadas (Jonckheere-Terpstra para amostras k) é uma alternativa mais eficaz para Kruskal-Wallis quando as amostras *k* têm uma ordem natural. Por exemplo, as populações *k* podem representar temperaturas aumentadas *k*. A hipótese de que temperaturas diferentes produzem a mesma distribuição de respostas é testada com relação à alternativa de que conforme a temperatura aumenta, a magnitude da resposta aumenta. Aqui, a hipótese alternativa é ordenada; portanto, Jonckheere-Terpstra é o teste mais apropriado para usar. **Menor para maior** especifica a hipótese alternativa de que o parâmetro de localização do primeiro grupo é menor que ou igual ao segundo, que é menor que ou igual ao terceiro, e assim por diante. **Maior para menor** especifica a hipótese alternativa de que o parâmetro de localização do primeiro grupo é maior que ou igual ao segundo, que é maior que ou igual ao terceiro, e assim por diante. Para ambas as opções, a hipótese alternativa também assume que as localizações não são iguais. É possível opcionalmente solicitar várias comparações das amostras *k*, todas as várias comparações **entre pares** ou comparações **de redução de Stepwise**.

- **Comparar Intervalos ao longo de Grupos.** Isso produz testes de amostras independentes de se as amostras têm o mesmo intervalo. **Reação extrema de Moisés (2 amostras)** testa um grupo de controle versus um grupo de comparação. O primeiro valor em ordem crescente do campo de agrupamento define o grupo de controle e o segundo define o grupo de comparação. Se o campo de agrupamento tiver mais de dois valores, esse teste não será produzido.
- **Comparar Medianas ao longo de Grupos.** Isso produz testes de amostras independentes de se as amostras têm a mesma mediana. **Teste de mediana (amostras k)** podem usar a mediana de amostra agrupada (calculada ao longo de todos os registros no conjunto de dados) ou um valor customizado como a mediana hipotética. É possível opcionalmente solicitar várias comparações das amostras *k*, todas as várias comparações **entre pares** ou comparações **de redução de Stepwise**.
- **Estimar Intervalos de Confiança ao longo de Grupos. A estimativa de Hodges-Lehman (2 amostras)** produz uma estimativa de amostras independentes e um intervalo de confiança para a diferença nas medianas de dois grupos. Se o campo de agrupamento tiver mais de dois valores, esse teste não será produzido.

Opções de Teste

Nível de Significância. Isso especifica o nível de significância (alpha) para todos os testes. Especifique um valor numérico entre 0 e 1. 0,05 é o padrão.

Intervalo de confiança (%). Isso especifica o nível de confiança para todos os intervalos de confiança produzidos. Especifique um valor numérico entre 0 e 100. 95 é o padrão.

Casos Excluídos. Isso especifica como determinar a base de caso para testes. **Excluir casos listwise** significa que registros com valores omissos para qualquer campo que é nomeado em qualquer subcomando são excluídos de todas as análises. **Excluir teste de casos por teste** significa que registros com valores omissos para um campo que é usado para um teste específico são omitidos desse teste. Quando diversos testes são especificados na análise, cada teste é avaliado separadamente.

Valores Omissos de Usuário

Valores Omissos de Usuário para Campos Categóricos. Campos categóricos devem ter valores válidos para um registro a ser incluído na análise. Esses controles permitem que você decida se os valores omissos de usuário são tratados como válidos entre os campos categóricos. Valores omissos do sistema e valores omissos para campos contínuos são sempre tratados como inválidos.

Variáveis Adicionais de Comando de NPTESTS

O idioma da sintaxe de comando também permite:

- Especificar testes de uma amostra, de amostras independentes e de amostras relacionadas em uma única execução do procedimento.

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Testes Não Paramétricos de Amostras Relacionadas

Identifica diferenças entre dois ou mais campos relacionados usando um ou mais testes não paramétricos. Os testes não paramétricos não presumem que seus dados sigam a distribuição normal.

Considerações de Dados. Cada registro corresponde a um determinado assunto para o qual duas ou mais medidas relacionadas estão armazenadas em campos separados no conjunto de dados. Por exemplo, um estudo referente à efetividade de um plano de dieta pode ser analisado usando testes não paramétricos de amostras relacionadas se cada ponderação do assunto for medida em intervalos regulares e armazenada em campos como *Peso pré-dieta*, *Peso intermediário* e *Peso pós-dieta*. Esses campos são "relacionados".

Qual é o seu objetivo? Os objetivos lhe permitem especificar rapidamente configurações de teste diferentes, mas usadas frequentemente.

- **Comparar automaticamente dados observados com dados hipotéticos.** Esse objetivo aplica Teste de McNemar em dados categóricos quando 2 campos são especificados, Q de Cochran em dados categóricos quando mais de 2 campos são especificados, o teste dos Postos Sinalizados de Pares Comparados de Wilcoxon para dados contínuos quando 2 campos são especificados e o ANOVA de 2 Vias de Friedman por Postos para dados contínuos quando mais de 2 campos são especificados.
- **Análise customizada.** Quando você desejar corrigir manualmente as configurações de teste na guia Configurações, selecione essa opção. Observe que essa configuração será selecionada automaticamente se você subsequentemente fizer mudanças em opções na guia Configurações que forem incompatíveis com o objetivo selecionado atualmente.

Quando campos de nível de medição diferente são especificados, eles são primeiro separados por nível de medição e, em seguida, o teste apropriado é aplicado a cada grupo. Por exemplo, se você escolher **Comparar automaticamente dados observados com dados hipotéticos** com o seu objetivo e especificar 3 campos contínuos e 2 campos nominais, então o teste de Friedman é aplicado nos campos contínuos e o teste de McNemar é aplicado nos campos nominais.

Para Obter Testes Não Paramétricos de Amostras Relacionadas

Nos menus, escolha:

Analisar > **Testes Não Paramétricos** > **Amostras Relacionadas...**

1. Clique em **Executar**.

Como opção, você pode:

- Especificar um objetivo na guia Objetivo.
- Especificar designações de campo na guia Campos.
- Especificar configurações de especialista na guia Configurações.

Guia Campos

A guia Campos especifica quais campos devem ser testados.

Usar funções predefinidas. Essa opção usa informação de campo. Todos os campos com uma função predefinida como Resposta ou Ambos serão usados como campos de teste. Pelo menos dois campos de teste são necessários.

Usar designações de campo customizado. Essa opção lhe permite substituir papéis do campo. Após selecionar essa opção, especifique os campos abaixo:

- **Campos de Teste.** Selecionar dois ou mais campos. Cada campo corresponde a uma amostra relacionada separada.

Guia Configurações

A guia Configurações abrange vários grupos diferentes de configurações que é possível modificar para ajustar com precisão como o procedimento processa os seus dados. Se você fizer qualquer mudança nas configurações padrão que for incompatível com os outros objetivos, a guia Objetivo será atualizada automaticamente para selecionar a opção **Customizar análise**.

Escolher Testes

Essas configurações especificam os testes a serem executados nos campos especificados na guia Campos.

Escolher automaticamente os testes com base nos dados. Essa configuração aplica Teste de McNemar em dados categóricos quando 2 campos são especificados, Q de Cochran em dados categóricos quando mais de 2 campos são especificados, o teste dos Postos Sinalizados de Pares Comparados de Wilcoxon para dados contínuos quando 2 campos são especificados e o ANOVA de 2 Vias de Friedman por Postos para dados contínuos quando mais de 2 campos são especificados.

Customizar testes. Essa configuração permite que você escolha testes específicos a serem executados.

- **Teste para Mudança em Dados Binários. Teste de McNemar (2 amostras)** pode ser aplicado em campos categóricos. Isso produz um teste de amostras relacionadas de se combinações de valores entre dois campos de sinalização (campos categóricos com apenas dois valores) são igualmente prováveis. Se houver mais do que dois campos especificados na guia Campos, esse teste não será realizado. Consulte "Teste de McNemar: Definir Sucesso" na página 135 para obter detalhes sobre as configurações de teste. **Q de Cochran (amostras k)** pode ser aplicado em campos categóricos. Isso produz um teste de amostras relacionadas de se combinações de valores entre campos de sinalização *k* (campos categóricos com apenas dois valores) são igualmente prováveis. É possível opcionalmente solicitar várias comparações das amostras *k*, todas as várias comparações **entre pares** ou comparações **de redução de stepwise**. Consulte "Q de Cochran: Definir Sucesso" na página 135 para obter detalhes sobre as configurações de teste.
- **Teste para Mudanças em Dados Multinomiais. Teste de homogeneidade marginal (2 amostras)** produz um teste de amostras relacionadas de se combinações de valores entre dois campos ordinais emparelhados são igualmente prováveis. O teste de homogeneidade marginal é normalmente usado em situações de medida repetida. Esse teste é uma extensão do teste de McNemar a partir da resposta binária até a resposta multinomial. Se houver mais do que dois campos especificados na guia Campos, esse teste não será realizado.
- **Comparar Diferença de Mediana com Hipotética.** Cada um desses testes produz um teste de amostras relacionadas de se a diferença de mediana entre dois campos é diferente de 0. O teste aplica-se a campos contínuos e ordinais. Se houver mais do que dois campos especificados na guia Campos, esses testes não serão realizados.
- **Estimar Intervalos de Confiança.** Isso produz uma estimativa de amostras relacionadas e um intervalo de confiança para a diferença de mediana entre dois campos emparelhados. O teste aplica-se a campos contínuos e ordinais. Se houver mais do que dois campos especificados na guia Campos, esse teste não será realizado.
- **Quantificar Associações. O coeficiente de concordância de Kendall (amostras k)** produz uma medida de concordância entre juízes ou avaliadores, em que cada registro é a classificação de diversos itens de um juiz (campos). É possível opcionalmente solicitar várias comparações das amostras *k*, todas as várias comparações **entre pares** ou comparações **de redução de Stepwise**.
- **Comparar Distribuições. ANOVA de 2 vias de Friedman por postos (amostras k)** produz um teste de amostras relacionadas de se amostras relacionadas *k* foram desenhadas a partir da mesma população. É possível opcionalmente solicitar várias comparações das amostras *k*, todas as várias comparações **entre pares** ou comparações **de redução de Stepwise**.

Teste de McNemar: Definir Sucesso: O teste de McNemar destina-se a campos de sinalização (campos categóricos com apenas duas categorias), mas é aplicado a todos os campos categóricos usando regras para definir "sucesso".

Definir Sucesso para Campos Categóricos. Isso especifica como "sucesso" é definido para campos categóricos.

- **Usar a primeira categoria localizada em dados** executa o teste usando o primeiro valor localizado na amostra para definir "sucesso". Essa opção é aplicável somente em campos nominais ou ordinais com apenas dois valores; todos os outros campos categóricos especificados na guia Campos em que essa opção é usada não serão testados. Este é o padrão.
- **Especificar valores de sucesso** executa o teste usando a lista especificada de valores para definir "sucesso". Especifique uma lista de sequências de caracteres ou de valores numéricos. Os valores na lista não precisam estar presentes na amostra.

Q de Cochran: Definir Sucesso: O teste Q de Cochran destina-se a campos de sinalização (campos categóricos com apenas duas categorias), mas é aplicado a todos os campos categóricos usando regras para definir "sucesso".

Definir Sucesso para Campos Categóricos. Isso especifica como "sucesso" é definido para campos categóricos.

- **Usar a primeira categoria localizada em dados** executa o teste usando o primeiro valor localizado na amostra para definir "sucesso". Essa opção é aplicável somente em campos nominais ou ordinais com apenas dois valores; todos os outros campos categóricos especificados na guia Campos em que essa opção é usada não serão testados. Esse é o padrão.
- **Especificar valores de sucesso** executa o teste usando a lista especificada de valores para definir "sucesso". Especifique uma lista de sequências de caracteres ou de valores numéricos. Os valores na lista não precisam estar presentes na amostra.

Opções de Teste

Nível de Significância. Isso especifica o nível de significância (alpha) para todos os testes. Especifique um valor numérico entre 0 e 1. 0,05 é o padrão.

Intervalo de confiança (%). Isso especifica o nível de confiança para todos os intervalos de confiança produzidos. Especifique um valor numérico entre 0 e 100. 95 é o padrão.

Casos Excluídos. Isso especifica como determinar a base de caso para testes.

- **Excluir casos listwise** significa que registros com valores omissos para qualquer campo que é nomeado em qualquer subcomando são excluídos de todas as análises.
- **Excluir teste de casos por teste** significa que registros com valores omissos para um campo que é usado para um teste específico são omitidos desse teste. Quando diversos testes são especificados na análise, cada teste é avaliado separadamente.

Valores Omissos de Usuário

Valores Omissos de Usuário para Campos Categóricos. Campos categóricos devem ter valores válidos para um registro a ser incluído na análise. Esses controles permitem que você decida se os valores omissos de usuário são tratados como válidos entre os campos categóricos. Valores omissos do sistema e valores omissos para campos contínuos são sempre tratados como inválidos.

Variáveis Adicionais de Comando de NPTESTS

O idioma da sintaxe de comando também permite:

- Especificar testes de uma amostra, de amostras independentes e de amostras relacionadas em uma única execução do procedimento.

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Visualização de Modelo

Visualização do Modelo

O procedimento cria um objeto Visualizador de Modelo no Visualizador. Ativando (clicando duas vezes em) esse objeto, você ganha uma visualização interativa do modelo. A visualização do modelo tem uma janela de dois painéis, a visualização principal à esquerda e a visualização vinculada ou auxiliar à direita.

Existem duas visualizações principais:

- Sumarização de Hipótese. Essa é a visualização padrão. Consulte o tópico "Sumarização de Hipótese" para obter mais informações.
- Sumarização de Intervalo de Confiança. Consulte o tópico "Sumarização do Intervalo de Confiança" para obter mais informações

Existem sete visualizações vinculadas/auxiliares:

- Teste de Uma Amostra. Essa é a visualização padrão se testes de uma amostra foram solicitados. Consulte o tópico "Teste de Uma Amostra" para obter mais informações
- Teste de Amostras Relacionadas. Essa é a visualização padrão se foram solicitados testes de amostras relacionadas e nenhum teste de uma amostra. Consulte o tópico "Teste de Amostras Relacionadas" na página 137 para obter mais informações
- Teste de Amostras Independentes. Essa é a visualização padrão se nenhum teste de amostras relacionadas ou testes de uma amostra foram solicitados. Consulte o tópico "Teste de Amostras Independentes" na página 138 para obter mais informações
- Informações de Campo Categórico. Consulte o tópico "Informação de Campo Categórico" na página 140 para obter mais informações
- Informações de Campo Contínuo. Consulte o tópico "Informação de Campo Contínuo" na página 140 para obter mais informações
- Comparações entre Pares. Consulte o tópico "Comparações Entre Pares" na página 140 para obter mais informações
- Subconjuntos Homogêneos. Consulte o tópico "Subconjuntos Homogêneos" na página 140 para obter mais informações

Sumarização de Hipótese

A visualização de Sumarização do Modelo é uma captura instantânea, sumarização de visão rápida dos testes não paramétricos. Ela enfatiza hipóteses e decisões nulas, chamando a atenção para valores *p* significativos.

- Cada linha corresponde a um teste separado. Clicar em uma linha mostra informações adicionais sobre o teste na visualização vinculada.
- Clicar em qualquer cabeçalho da coluna ordena as linhas por valores nessa coluna.
- O botão **Reconfigurar** permite que você retorne o Visualizador de Modelo para o seu estado original.
- A lista suspensa de **Filtros de Campo** permite que você exiba apenas os testes que envolvem o campo selecionado.

Sumarização do Intervalo de Confiança

A Sumarização do Intervalo de Confiança mostra qualquer intervalo de confiança produzido pelos testes não paramétricos.

- Cada linha corresponde a um intervalo de confiança separado.
- Clicar em qualquer cabeçalho da coluna ordena as linhas por valores nessa coluna.

Teste de Uma Amostra

A visualização de Teste de Uma Amostra mostra detalhes relacionados a qualquer teste não paramétrico de uma amostra. As informações mostradas dependem do teste selecionado.

- A lista suspensa de **Testes** lhe permite selecionar um determinado tipo de teste de uma amostra.
- A lista suspensa de **Campos** lhe permite selecionar um campo que foi testado usando o teste selecionado na lista suspensa de **Testes**.

Teste de Binômio

O Teste de Binômio mostra um gráfico de barras empilhadas e uma tabela de teste.

- O gráfico de barras empilhadas exibe as frequências observadas e hipotéticas para as categorias "sucesso" e "falha" do campo de teste, com "falhas" empilhadas na parte superior de "sucessos". Passar o mouse sobre uma barra mostra as porcentagens de categoria em uma dica de ferramenta. Diferenças visíveis nas barras indicam que o campo de teste pode não ter a distribuição binomial hipotética.
- A tabela mostra detalhes do teste.

Teste qui-quadrado

A visualização de Teste Qui-Quadrado mostra um gráfico de barras agrupadas e uma tabela de teste.

- O gráfico de barras agrupadas exibe as frequências observadas e hipotéticas para cada categoria do campo de teste. Passar o mouse sobre uma barra mostra as frequências observadas e hipotéticas e sua diferença (de resíduo) em uma dica de ferramenta. Diferenças visíveis nas barras observadas versus hipotéticas indicam que o campo de teste pode não ter a distribuição hipotética.
- A tabela mostra detalhes do teste.

Postos Sinalizados de Wilcoxon

A visualização de Teste dos Postos Sinalizados de Wilcoxon mostra um histograma e uma tabela de teste.

- O histograma inclui linhas verticais mostrando as medianas observadas e hipotéticas.
- A tabela mostra detalhes do teste.

Teste de Execuções

A visualização de Teste de Execuções mostra um gráfico e uma tabela de teste.

- O gráfico exibe uma distribuição normal com o número observado de execuções marcado com uma linha vertical. Observe que quando o teste exato é executado, o teste não é baseado na distribuição normal.
- A tabela mostra detalhes do teste.

Teste de Kolmogorov-Smirnov

A visualização de Teste de Kolmogorov-Smirnov mostra um histograma e uma tabela de teste.

- O histograma inclui uma sobreposição da função de densidade de probabilidade para a distribuição uniforme hipotética, normal, de Poisson ou exponencial. Observe que o teste é baseado em distribuições acumulativas e as Diferenças Mais Extremas relatadas na tabela devem ser interpretadas com relação a distribuições acumulativas.
- A tabela mostra detalhes do teste.

Teste de Amostras Relacionadas

A visualização de Teste de Uma Amostra mostra detalhes relacionados a qualquer teste não paramétrico de uma amostra. As informações mostradas dependem do teste selecionado.

- A lista suspensa de **Testes** lhe permite selecionar um determinado tipo de teste de uma amostra.
- A lista suspensa de **Campos** lhe permite selecionar um campo que foi testado usando o teste selecionado na lista suspensa de **Testes**.

Teste de McNemar

A visualização de Teste de McNemar mostra um gráfico de barras agrupadas e uma tabela de teste.

- O gráfico de barras agrupadas exibe as frequências observadas e hipotéticas para as células fora da diagonal da tabela 2 × 2 definida pelos campos de teste.
- A tabela mostra detalhes do teste.

Teste de Sinal

A visualização de Teste de Sinal mostra um histograma empilhado e uma tabela de teste.

- O histograma empilhado exibe as diferenças entre os campos, usando o sinal da diferença como o campo de empilhamento.
- A tabela mostra detalhes do teste.

Teste dos Postos Sinalizados de Wilcoxon

A visualização de Teste dos Postos Sinalizados de Wilcoxon mostra um histograma empilhado e uma tabela de teste.

- O histograma empilhado exibe as diferenças entre os campos, usando o sinal da diferença como o campo de empilhamento.
- A tabela mostra detalhes do teste.

Teste de Homogeneidade Marginal

A visualização de Teste de Homogeneidade Marginal mostra um gráfico de barras agrupadas e uma tabela de teste.

- O gráfico de barras agrupadas exibe as frequências observadas para as células fora da diagonal da tabela definida pelos campos de teste.
- A tabela mostra detalhes do teste.

Teste Q de Cochran

A visualização de Teste Q de Cochran mostra um gráfico de barras empilhadas e uma tabela de teste.

- O gráfico de barras empilhadas exibe as frequências observadas para as categorias "sucesso" e "falha" dos campos de teste, com "falhas" empilhadas na parte superior de "sucessos". Passar o mouse sobre uma barra mostra as porcentagens de categoria em uma dica de ferramenta.
- A tabela mostra detalhes do teste.

Análise de Variância de Dois Fatores de Friedman por Postos

A visualização de Análise de Variância de Dois Fatores de Friedman por Postos mostra histogramas em painel e uma tabela de teste.

- Os histogramas exibem a distribuição observada de postos, colocados em painel pelos campos de teste.
- A tabela mostra detalhes do teste.

Coeficiente de Concordância de Kendall

A visualização de Coeficiente de Concordância de Kendall mostra histogramas em painel e uma tabela de teste.

- Os histogramas exibem a distribuição observada de postos, colocados em painel pelos campos de teste.
- A tabela mostra detalhes do teste.

Teste de Amostras Independentes

A visualização de Teste de Amostras Independentes mostra detalhes relacionados a qualquer teste não paramétrico de amostras independentes. As informações mostradas dependem do teste selecionado.

- A lista suspensa de **Testes** lhe permite selecionar um determinado tipo de teste de amostras independentes.
- A lista suspensa de **Campos** lhe permite selecionar uma combinação de campos de teste e agrupamento que foi testada usando o teste selecionado na lista suspensa de **Testes**.

Teste de Mann-Whitney

A visualização de Teste de Mann-Whitney mostra um gráfico de pirâmide populacional e uma tabela de teste.

- O gráfico de pirâmide populacional exibe histogramas back-to-back pelas categorias do campo de agrupamento, anotando o número de registros em cada grupo e a classificação média do grupo.
- A tabela mostra detalhes do teste.

Teste de Kolmogorov-Smirnov

A visualização de Teste de Kolmogorov-Smirnov mostra um gráfico de pirâmide populacional e uma tabela de teste.

- O gráfico de pirâmide populacional exibe histogramas back-to-back pelas categorias do campo de agrupamento, anotando o número de registros em cada grupo. As linhas de distribuição acumulativas observadas podem ser exibidas ou ocultadas clicando no botão **Acumulativo**.
- A tabela mostra detalhes do teste.

Teste de Execuções de Wald-Wolfowitz

A visualização de Teste de Execuções de Wald-Wolfowitz mostra um gráfico de barras empilhadas e uma tabela de teste.

- O gráfico de pirâmide populacional exibe histogramas back-to-back pelas categorias do campo de agrupamento, anotando o número de registros em cada grupo.
- A tabela mostra detalhes do teste.

Teste de Kruskal-Wallis

A visualização de Teste de Kruskal-Wallis mostra boxplots e uma tabela de teste.

- Boxplots separados são exibidos para cada categoria do campo de agrupamento. Passar o mouse sobre uma caixa mostra a classificação média em uma dica de ferramenta.
- A tabela mostra detalhes do teste.

Teste de Jonckheere-Terpstra

A visualização de Teste de Jonckheere-Terpstra mostra boxplots e uma tabela de teste.

- Boxplots separados são exibidos para cada categoria do campo de agrupamento.
- A tabela mostra detalhes do teste.

Teste de Moisés de Reação Extrema

A visualização de Teste de Moisés de Reação Extrema mostra boxplots e uma tabela de teste.

- Boxplots separados são exibidos para cada categoria do campo de agrupamento. Os rótulos de ponto podem ser exibidos ou ocultados clicando no botão **ID de Registro**.
- A tabela mostra detalhes do teste.

Teste de Mediana

A visualização de Teste de Mediana mostra box plots e uma tabela de teste.

- Boxplots separados são exibidos para cada categoria do campo de agrupamento.
- A tabela mostra detalhes do teste.

Informação de Campo Categórico

A visualização de Informação de Campo Categórico exibe um gráfico de barras para o campo categórico selecionado na lista suspensa de **Campos**. A lista de campos disponíveis é restrita aos campos categóricos usados no teste selecionado no momento na visualização de Sumarização de Hipótese.

- Passar o mouse sobre uma barra fornece à categoria porcentagens em uma dica de ferramenta.

Informação de Campo Contínuo

A visualização de Informação de Campo Contínuo exibe um histograma para o campo contínuo selecionado na lista suspensa de **Campos**. A lista de campos disponíveis é restrita aos campos contínuos usados no teste selecionado no momento na visualização de Sumarização de Hipótese.

Comparações Entre Pares

A visualização de Comparações Entre Pares mostra um gráfico de rede de distância e a tabela de comparações produzida por testes não paramétricos de amostra *k* quando várias comparações entre pares são solicitadas.

- O gráfico de rede de distância é uma representação gráfica da tabela de comparações na qual as distâncias entre nós na rede correspondem a diferenças entre amostras. Linhas amarelas correspondem a diferenças estatisticamente significativas; linhas pretas correspondem a diferenças não significativas. Passar o mouse sobre uma linha na rede exibe uma dica de ferramenta com a significância ajustada da diferença entre os nós conectados pela linha.
- A tabela de comparação mostra os resultados numéricos de todas as comparações entre pares. Cada linha corresponde a uma comparação entre pares separada. Clicar em um cabeçalho da coluna ordena as linhas por valores nessa coluna.

Subconjuntos Homogêneos

A visualização dos Subconjuntos Homogêneos mostra uma tabela de comparações produzida por testes não paramétricos de amostra *k* quando várias comparações de redução de stepwise são solicitadas.

- Cada linha no Grupo de amostra corresponde a uma amostra relacionada separada (representada nos dados por campos separados). Amostras que não são estatística e significantemente diferentes são agrupadas nos subconjuntos da mesma cor; há uma coluna separada para cada subconjunto identificado. Quando todas as amostras são estatística e significantemente diferentes, há um subconjunto separado para cada amostra. Quando nenhuma das amostras é estatística e significantemente diferente, há um subconjunto único.
- Uma estatística do teste, um valor de significância e um valor de significância ajustado são calculados para cada subconjunto que contém mais de uma amostra.

Variáveis Adicionais de Comando de NPTESTS

O idioma da sintaxe de comando também permite:

- Especificar testes de uma amostra, de amostras independentes e de amostras relacionadas em uma única execução do procedimento.

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Diálogos Anteriores

Há uma série de diálogos "anteriores" que também executam testes não paramétricos. Esses diálogos suportam a funcionalidade fornecida pela opção de Testes Exatos.

Teste Qui-Quadrado. Tabula uma variável em categorias e calcula uma estatística qui-quadrado baseada nas diferenças entre frequências observadas e esperadas.

Teste Binomial. Compara a frequência observada em cada categoria de uma variável dicotômica com frequências esperadas a partir da distribuição binomial.

Teste de Execuções. Testa se a ordem de ocorrência de dois valores de uma variável é aleatória.

Teste de Kolmogorov-Smirnov de Uma Amostra. Compara a função de distribuição acumulativa observada para uma variável com uma distribuição teórica especificada, que pode ser normal, uniforme, exponencial ou de Poisson.

Testes de Duas Amostras Independentes. Compara dois grupos de casos em uma variável. O teste *U* de Mann-Whitney, o teste de Kolmogorov-Smirnov de duas amostras, o teste de Moisés de reações extremas e o teste de sequências de Wald-Wolfowitz estão disponíveis.

Testes de Duas Amostras Relacionadas. Compara as distribuições de duas variáveis. O teste dos postos sinalizados de Wilcoxon, o teste de sinal e o teste de McNemar estão disponíveis.

Testes para Várias Amostras Independentes. Compara dois ou mais grupos de casos em uma variável. O teste de Kruskal-Wallis, o teste de Mediana e o teste de Jonckheere-Terpstra estão disponíveis.

Testes para Várias Amostras Relacionadas. Compara as distribuições de duas ou mais variáveis. O teste de Friedman, *W* de Kendall e *Q* de Cochran estão disponíveis.

Quartis e a média, desvio padrão, mínimo, máximo e número de casos não omissos estão disponíveis para todos os testes acima.

Teste qui-quadrado

O procedimento Teste Qui-Quadrado tabula uma variável em categorias e calcula uma estatística qui-quadrado. Este teste de Qualidade de ajuste compara as frequências observadas e esperadas em cada categoria para confirmar que todas as categorias contêm a mesma proporção de valores ou confirmar que cada categoria contém uma proporção de valores especificada pelo usuário.

Exemplos. O teste qui-quadrado pode ser utilizado para determinar se um pacote de balas contém proporções iguais de balas azuis, marrons, verdes, laranjas, vermelhas e amarelas. Também é possível testar se um saco de balas contém 5% de balas azuis, 30% de marrons, 10% de verdes, 20% de laranjas, 15% de vermelhas e 15% de amarelas.

Estatísticas. Média, desvio padrão, mínimo, máximo, e quartis. O número e a porcentagem de casos não omissos e omissos, o número de casos observados e esperados para cada categoria, resíduos e estatística qui-quadrado.

Considerações de Dados de Teste Qui-Quadrado

Dados. Utilize variáveis categóricas numéricas ordenadas ou não ordenadas (níveis de medição ordinais ou nominais). Para converter variáveis de sequência de caracteres em variáveis numéricas, utilize o procedimento Recodificação Automática, que está disponível no menu Transformação.

Suposições. Testes não paramétricos não requerem suposições sobre a forma da distribuição subjacente. Os dados são considerados como uma amostra aleatória. As frequências esperadas para cada categoria devem ser pelo menos 1. Não mais que 20% das categorias devem ter frequências esperadas inferiores a 5.

Para Obter um Teste Qui-Quadrado

1. Nos menus, escolha:

   Analisar > **Testes Não Paramétricos** > **Diálogos Anteriores** > **Qui-Quadrado...**
2. Selecione uma ou mais variáveis de teste. Cada variável produz um teste separado.

3. Opcionalmente, clique em **Opções** para estatísticas descritivas, quartis e controle do tratamento de dados omissos.

Intervalo e Valores Esperados de Teste Qui-Quadrado

Intervalo Esperado. Por padrão, cada valor distinto da variável é definido como uma categoria. Para estabelecer categorias dentro de um intervalo específico, selecione **Usar intervalo especificado** e insira valores de número inteiro para limites inferior e superior. Categorias são estabelecidas para cada valor de número inteiro dentro do intervalo inclusivo, e os casos com valores fora dos limites são excluídos. Por exemplo, se especificar um valor de 1 para Inferior e um valor de 4 para Superior, apenas os valores de número inteiro de 1 a 4 serão utilizados para o teste qui-quadrado.

Valores Esperados. Por padrão, todas as categorias possuem valores iguais esperados. As categorias podem ter proporções esperadas especificadas pelo usuário. Selecione **Valores**, insira um valor que seja maior que 0 para cada categoria da variável de teste e, em seguida, clique em **Incluir**. Sempre que um valor é incluído, ele aparece no final da lista de valores. A ordem dos valores é importante porque ela corresponde à ordem crescente dos valores da categoria da variável de teste. O primeiro valor na lista corresponde ao menor valor do grupo da variável de teste, e o último valor corresponde ao valor mais alto. Os elementos da lista de valores são somados e, em seguida, cada valor é dividido por esta soma para calcular a proporção de casos esperados na categoria correspondente. Por exemplo, uma lista de valores de 3, 4, 5, 4 especifica proporções esperadas de 3/16, 4/16, 5/16, e 4/16.

Opções de Teste Qui-Quadrado

Estatísticas. É possível escolher uma ou ambas as estatísticas básicas.

- **Descritivo.** Exibe a média, o desvio padrão, o mínimo, o máximo e o número de casos não desconhecidos.
- **Quartis.** Exibe os valores correspondentes aos 25⌂, 50⌂ e 75⌂ percentis.

Valores omissos. Controla o tratamento de valores omissos.

- **Excluir casos de teste por teste.** Quando vários testes são especificados, cada teste é avaliado separadamente para valores omissos.
- **Excluir listwise dos casos.** Casos com valores omissos para qualquer variável são excluídos de todas as análises.

Recursos Adicionais do Comando NPAR TESTS (Teste Qui-Quadrado)

O idioma da sintaxe de comando também permite:

- Especificar valores mínimo e máximo ou frequências esperadas diferentes para variáveis diferentes (com o subcomando `CHISQUARE`).
- Testar a mesma variável em diferentes frequências esperadas ou utilizar intervalos diferentes (com o subcomando `EXPECTED`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Teste de Binômio

O procedimento de Teste Binomial compara as frequências observadas das duas categorias de uma variável dicotômica com as frequências que são esperadas em uma distribuição binomial com um parâmetro de probabilidade especificado. Por padrão, o parâmetro de probabilidade para ambos os grupos é 0,5. Para alterar as probabilidades, é possível inserir uma proporção de teste para o primeiro grupo. A probabilidade para o segundo grupo será 1 menos a probabilidade especificada para o primeiro grupo.

Exemplo. Quando você joga uma moeda, a probabilidade de cara igual a 1/2. Com base nesta hipótese, uma moeda é lançada 40 vezes, e os resultados são registrados (cara ou coroa). No teste Binomial, talvez

você ache que 3/4 dos lançamentos eram caras e que o nível de significância observado é pequeno (0,0027). Esses resultados indicam que não é provável que a probabilidade de cara seja de 1/2; a moeda provavelmente é tendenciosa.

Estatísticas. Média, desvio padrão, mínimo, máximo, número de casos não desconhecidos e quartis.

Considerações de Dados de Teste Binomial

Dados. As variáveis que são testadas devem ser numéricas e dicotômicas. Para converter variáveis de sequência de caracteres em variáveis numéricas, utilize o procedimento Recodificação Automática, que está disponível no menu Transformação. Uma **variável dicotômica** é uma variável que pode ter apenas dois valores possíveis: *sim* ou *não*, *true* ou *false*, 0 ou 1, e assim por diante. O primeiro valor encontrado no conjunto de dados define o primeiro grupo, e o outro valor define o segundo grupo. Se as variáveis não forem dicotômicas, deve-se especificar um ponto de corte. O ponto de corte designa casos com valores que forem menores ou iguais ao ponto de corte para o primeiro grupo e designa o restante dos casos para o segundo grupo.

Suposições. Testes não paramétricos não requerem suposições sobre a forma da distribuição subjacente. Os dados são considerados como uma amostra aleatória.

Para Obter um Teste Binomial

1. Nos menus, escolha:

   Analisar > **Testes Não Paramétricos** > **Diálogos Anteriores** > **Binomial...**
2. Selecione uma ou mais variáveis de teste numéricas.
3. Opcionalmente, clique em **Opções** para estatísticas descritivas, quartis e controle do tratamento de dados omissos.

Opções de Teste Binomial

Estatísticas. É possível escolher uma ou ambas as estatísticas básicas.

- **Descritivo.** Exibe a média, o desvio padrão, o mínimo, o máximo e o número de casos não desconhecidos.
- **Quartis.** Exibe os valores correspondentes aos 25⌂, 50⌂ e 75⌂ percentis.

Valores omissos. Controla o tratamento de valores omissos.

- **Excluir casos de teste por teste.** Quando vários testes são especificados, cada teste é avaliado separadamente para valores omissos.
- **Excluir listwise dos casos.** Casos com valores omissos para qualquer variável que é testada são excluídos de todas as análises.

Recursos Adicionais do Comando NPAR TESTS (Teste Binomial)

O idioma da sintaxe de comando também permite:

- Selecionar grupos específicos (e excluir outros grupos) quando uma variável tiver mais de duas categorias (com o subcomando `BINOMIAL`).
- Especificar diferentes pontos de corte ou probabilidades para varáveis diferentes (com o subcomando `BINOMIAL`).
- Testar a mesma variável em diferentes pontos de corte ou probabilidades (com o subcomando `EXPECTED`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Teste de Execuções

O procedimento Teste de Execuções testa se a ordem de ocorrência de dois valores de uma variável é aleatória. Uma execução é uma sequência de observações semelhantes. Uma amostra com muitas ou poucas execuções sugere que a amostra não é aleatória.

Exemplos. Suponha que 20 pessoas sejam reunidas para descobrir se elas comprariam um produto. A aleatoriedade suposta da amostra seria questionada seriamente se todas as 20 pessoas fossem do mesmo sexo. O teste de execuções pode ser utilizado para determinar se a amostra foi tirada aleatoriamente.

Estatísticas. Média, desvio padrão, mínimo, máximo, número de casos não desconhecidos e quartis.

Considerações de Dados de Teste de Execuções

Dados. As variáveis devem ser numéricas. Para converter variáveis de sequência de caracteres em variáveis numéricas, utilize o procedimento Recodificação Automática, que está disponível no menu Transformação.

Suposições. Testes não paramétricos não requerem suposições sobre a forma da distribuição subjacente. Use amostras de distribuições de probabilidade contínuas.

Para Obter um Teste de Execuções

1. Nos menus, escolha:

   Analisar > **Testes Não Paramétricos** > **Diálogos Anteriores** > **Execuções...**
2. Selecione uma ou mais variáveis de teste numéricas.
3. Opcionalmente, clique em **Opções** para estatísticas descritivas, quartis e controle do tratamento de dados omissos.

Ponto de Corte de Teste de Execuções

Ponto de corte. Especifica um ponto de corte para dicotomizar as variáveis que você escolheu. É possível utilizar a média, a mediana ou o modo observado ou utilizar um valor especificado como um ponto de corte. Casos com valores menores que o ponto de corte são designados a um grupo, e os casos com valores maiores que ou iguais ao ponto de corte são designados para outro grupo. Um teste é executado para cada ponto de corte escolhido.

Opções de Teste de Execuções

Estatísticas. É possível escolher uma ou ambas as estatísticas básicas.

- **Descritivo.** Exibe a média, o desvio padrão, o mínimo, o máximo e o número de casos não desconhecidos.
- **Quartis.** Exibe os valores correspondentes aos 25⌂, 50⌂ e 75⌂ percentis.

Valores omissos. Controla o tratamento de valores omissos.

- **Excluir casos de teste por teste.** Quando vários testes são especificados, cada teste é avaliado separadamente para valores omissos.
- **Excluir listwise dos casos.** Casos com valores omissos para qualquer variável são excluídos de todas as análises.

Recursos Adicionais do Comando NPAR TESTS (Teste de Execuções)

O idioma da sintaxe de comando também permite:

- Especificar diferentes pontos de corte para varáveis diferentes (com o subcomando RUNS).
- Testar a mesma variável em diferentes pontos de corte customizados (com o subcomando RUNS).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Teste de Kolmogorov-Smirnov de uma amostra

O procedimento Teste de Kolmogorov-Smirnov de Uma Amostra compara a função de distribuição cumulativa observada de uma variável com uma distribuição teórica especificada, que pode ser normal, uniforme, Poisson ou exponencial. O Z de Kolmogorov-Smirnov é calculado a partir da diferença maior (em valor absoluto) entre as funções de distribuição cumulativa observadas e teóricas. Esse teste de Qualidade do ajuste testa se as observações podem razoavelmente ter vindo da distribuição especificada.

Exemplo. Muitos testes paramétricos requerem variáveis normalmente distribuídas. O teste de Kolmogorov-Smirnov de uma amostra pode ser utilizado para testar se uma variável (por exemplo, *income*) é normalmente distribuída.

Estatísticas. Média, desvio padrão, mínimo, máximo, número de casos não desconhecidos e quartis.

Considerações de Dados do Teste de Kolmogorov-Smirnov de Uma Amostra

Dados. Utilize variáveis quantitativas (nível de medição de intervalo ou razão).

Suposições. O teste de Kolmogorov-Smirnov supõe que os parâmetros da distribuição de teste são especificados antecipadamente. Este procedimento estima os parâmetros a partir da amostra. A média da amostra e o desvio padrão de amostra são os parâmetros para uma distribuição normal, os valores mínimo e máximo da amostra definem o intervalo da distribuição uniforme, a média da amostra é o parâmetro para a distribuição de Poisson, e a média da amostra é o parâmetro para a distribuição exponencial. O poder do teste para detectar partidas da distribuição hipotética pode ser seriamente prejudicado. Para teste com relação a uma distribuição normal com os parâmetros estimados, considere o teste K-S Lilliefors ajustado (disponível no procedimento Explorar).

Para Obter um Teste de Kolmogorov-Smirnov de Uma Amostra

1. Nos menus, escolha:

   Analisar > **Testes Não Paramétricos** > **Diálogos Anteriores** > **K-S de 1 Amostra...**
2. Selecione uma ou mais variáveis de teste numéricas. Cada variável produz um teste separado.
3. Opcionalmente, clique em **Opções** para estatísticas descritivas, quartis e controle do tratamento de dados omissos.

Opções de Teste de Kolmogorov-Smirnov de Uma Amostra

Estatísticas. É possível escolher uma ou ambas as estatísticas básicas.

- **Descritivo.** Exibe a média, o desvio padrão, o mínimo, o máximo e o número de casos não desconhecidos.
- **Quartis.** Exibe os valores correspondentes aos 25⌂, 50⌂ e 75⌂ percentis.

Valores omissos. Controla o tratamento de valores omissos.

- **Excluir casos de teste por teste.** Quando vários testes são especificados, cada teste é avaliado separadamente para valores omissos.
- **Excluir listwise dos casos.** Casos com valores omissos para qualquer variável são excluídos de todas as análises.

Recursos Adicionais do Comando NPAR TESTS (Teste de Kolmogorov-Smirnov de Uma Amostra)

A linguagem de sintaxe de comando também permite especificar os parâmetros da distribuição de teste (com o subcomando `K-S`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Testes de duas amostras independentes

O procedimento Testes de Duas Amostras Independentes compara dois grupos de casos em uma variável.

Exemplo. Novos aparelhos dentários foram desenvolvidos para que sejam mais confortáveis, tenham uma melhor aparência e forneçam progresso mais rápido no realinhamento dos dentes. Para descobrir se os novos aparelhos foram usados tanto quanto os aparelhos antigos, 10 crianças foram escolhidas aleatoriamente para usar aparelhos antigos, e outras 10 crianças foram escolhidas para usar aparelhos novos. No teste *U* de Mann-Whitney, você pode descobrir que, em média, as crianças com os novos aparelhos não precisaram utilizá-los tanto quanto os aparelhos antigos.

Estatísticas. Média, desvio padrão, mínimo, máximo, número de casos não desconhecidos e quartis. Testes: *U* de Mann-Whitney, reações extremas de Moses, *Z* de Kolmogorov-Smirnov, execuções de Wald-Wolfowitz.

Considerações de Dados de Testes de Duas Amostras Independentes

Dados. Utilize variáveis numéricas que possam ser ordenadas.

Suposições. Utilize amostras aleatórias independentes. O teste *U* de Mann-Whitney testa a igualdade de duas distribuições. Para que isso seja usado para testar as diferenças na localização entre duas distribuições, alguém deve supor que as distribuições têm a mesma forma.

Para Obter Testes de Duas Amostras Independentes

1. Nos menus, escolha:

   Analisar > **Testes Não Paramétricos** > **Diálogos Anteriores** > **2 Amostras Independentes ...**
2. Selecione uma ou mais variáveis numéricas.
3. Selecione uma variável de agrupamento e clique em **Definir Grupos** para dividir o arquivo em dois grupos ou amostras.

Tipos de teste Duas amostras independentes

Tipo de teste. Quatro testes estão disponíveis para testar se duas amostras independentes (grupos) vêm da mesma população.

O **teste U de Mann-Whitney** é o mais popular dos testes de duas amostras independentes. Ele é equivalente ao teste de soma de ranqueamento Wilcoxon e ao teste de Kruskal-Wallis para dois grupos. O Mann-Whitney testa que duas populações amostradas são equivalentes em localização. As observações de ambos os grupos são combinadas e classificadas, com o ranqueamento médio designado no caso de empates. O número de empates deve ser pequeno em relação ao número total de observações. Se as populações são idênticas em localização, as classificações devem ser aleatoriamente combinadas entre as duas amostras. O teste calcula o número de vezes que uma pontuação do 1 grupo precede uma pontuação do grupo 2 e o número de vezes que uma pontuação do 2 grupo precede uma pontuação do grupo 1. A estatística do Mann-Whitney *U* é a menor dos dois números. A estatística *W* da soma de ranqueamento do Wilcoxon também é exibida. *W* é a soma das classificações para o grupo com a menor classificação média, a menos que os grupos tenham a mesma classificação média, nestes caso, é a soma da classificação do grupo que é nomeado por último na caixa de diálogo Grupos de definição de duas amostras independentes.

O **teste Z de Kolmogorov-Smirnov** e o **teste de sequências de Wald-Wolfowitz** são testes mais gerais que detectam diferenças em ambas as localizações e formas das distribuições. O teste de Kolmogorov-Smirnov baseia-se na diferença absoluta máxima entre as funções de distribuição acumulativas observadas para ambas as amostras. Quando essa diferença é significativamente grande, as duas distribuições são consideradas diferentes. O teste de sequências de Wald-Wolfowitz combina e classifica as observações de ambos os grupos. Se as duas amostras forem da mesma população, os dois grupos deverão ser dispersos aleatoriamente em todo o ranqueamento.

O **teste de reações extremas de Moses** assume que a variável experimental afetará alguns assuntos em uma direção e outros assuntos na direção oposta. Os testes testam para obter respostas extremas em comparação aum grupo de controle. Este teste foca no span do grupo de controle e é uma medida do quanto os valores extremos no grupo experimental influenciam o span quando combinados com o grupo de controle. O grupo de controle é definido pelo valor do grupo 1 na caixa de diálogo Grupos de definição de duas amostras independentes. Observações de ambos os grupos são combinadas e classificadas. O span do grupo de controle é calculado como a diferença entre as classificações dos valores mais altos e mais baxiso no grupo de controle mais 1. Como valores discrepantes de chances podem distorcer facilmente o intervalo do span, 5% dos casos de controle são cortados automaticamente de cada extremidade.

Definir Grupos para Testes de Duas Amostras Independentes

Para dividir o arquivo em dois grupos ou amostras, insira um valor de número inteiro para o Grupo 1 e outro valor para o Grupo 2. Casos com outros valores são excluídos da análise.

Opções de Testes de Duas Amostras Independentes

Estatísticas. É possível escolher uma ou ambas as estatísticas básicas.

- **Descritivo.** Exibe a média, o desvio padrão, o mínimo, o máximo e o número de casos não desconhecidos.
- **Quartis.** Exibe os valores correspondentes aos 25⌂, 50⌂ e 75⌂ percentis.

Valores omissos. Controla o tratamento de valores omissos.

- **Excluir casos de teste por teste.** Quando vários testes são especificados, cada teste é avaliado separadamente para valores omissos.
- **Excluir listwise dos casos.** Casos com valores omissos para qualquer variável são excluídos de todas as análises.

Recursos Adicionais do Comando NPAR TESTS (Testes de Duas Amostras Independentes)

A linguagem de sintaxe de comando também permite especificar o número de casos a serem aparados para o teste de Moses (com o subcomando `MOSES`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Testes de Duas Amostras Relacionadas

O procedimento Testes de Duas Amostras Relacionadas compara as distribuições de duas variáveis.

Exemplo. Em geral, as famílias recebem o preço que pediram quando elas vendem suas casas? Ao aplicar o teste de postos sinalizados de Wilcoxon para dados de 10 residências, você poderá saber que sete famílias recebem menos do que o preço pedido, uma família recebe mais do que o preço pedido, e duas famílias recebem o preço pedido.

Estatísticas. Média, desvio padrão, mínimo, máximo, número de casos não desconhecidos e quartis. Testes: postos sinalizados de Wilcoxon, assinar, McNemar. Se a opção Testes Exatos estiver instalada (disponível apenas em sistemas operacionais Windows), o teste de homogeneidade marginal também está disponível.

Considerações de Dados de Testes de Duas Amostras Relacionadas

Dados. Utilize variáveis numéricas que possam ser ordenadas.

Suposições. Embora nenhuma distribuição específica seja assumida para as duas variáveis, a distribuição da população das diferenças pairwise é assumida como simétrica.

Para Obter Testes de Duas Amostras Relacionadas

1. Nos menus, escolha:

   Analisar > **Testes Não Paramétricos** > **Diálogos Anteriores** > **2 Amostras Relacionadas ...**
2. Selecione um ou mais pares de variáveis.

Tipos de teste de duas amostras relacionadas

Os testes desta seção comparam as distribuições de duas variáveis relacionadas. O teste apropriado para uso depende do tipo de dados.

Se seus dados são contínuos, use o teste de sinal ou o teste dos postos sinalizados de Wilcoxon. O **teste de sinal** calcula as diferenças entre as duas variáveis para todos os casos e classifica os diferenças como positivas, negativas ou empatadas. Se as duas variáveis forem igualmente distribuídas, o número de diferenças positivas e negativas não diferirá significativamente. O **teste dos postos sinalizados de Wilcoxon** considera informações sobre o sinal das diferenças e a magnitude das diferenças entre os pares. Como o teste dos postos sinalizados de Wilcoxon incorpora mais informações sobre os dados, ele é mais poderoso que o teste de sinal.

Se seus dados são binários, use o **teste de McNemar**. Este teste é normalmente usado em uma situação de medidas repetidas, em que cada resposta do assunto é provocada duas vezes, uma vez antes e uma vez após um evento especificado ocorrer. O teste de McNemar determina se a taxa de resposta inicial (antes do evento) é igual à taxa de resposta final (após o evento). Este teste é útil para detectar mudanças nas respostas devido à intervenção experimental em designs de antes e depois.

Se seus dados são categóricos, use o **teste de homogeneidade marginal**. Este teste é uma extensão do teste de McNemar da resposta binária com a resposta multinomial. Ele testa para obter mudanças nas respostas (usando a distribuição qui-quadrado) e é útil para detectar mudanças de resposta devido à intervenção experimental em designs de antes e depois. O teste de homogeneidade marginal está disponível apenas se você tiver instalado Testes Exatos.

Opções de Testes de Duas Amostras Relacionadas

Estatísticas. É possível escolher uma ou ambas as estatísticas básicas.

- **Descritivo.** Exibe a média, o desvio padrão, o mínimo, o máximo e o número de casos não desconhecidos.
- **Quartis.** Exibe os valores correspondentes aos 25⌂, 50⌂ e 75⌂ percentis.

Valores omissos. Controla o tratamento de valores omissos.

- **Excluir casos de teste por teste.** Quando vários testes são especificados, cada teste é avaliado separadamente para valores omissos.
- **Excluir listwise dos casos.** Casos com valores omissos para qualquer variável são excluídos de todas as análises.

Recursos Adicionais do Comando NPAR TESTS (Teste Qui-Quadrado)

A linguagem de sintaxe de comando também permite testar uma variável com cada variável em uma lista.

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Testes para diversas amostras independentes

O procedimento Testes para Várias Amostras Independentes compara dois ou mais grupos de casos em uma variável.

Exemplo. Três marcas de lâmpadas de 100 watts diferem no tempo médio de vida em que elas irão queimar? Na análise de variância para um fator de Kruskal-Wallis, você pode ver que as três marcas diferem no tempo médio de vida.

Estatísticas. Média, desvio padrão, mínimo, máximo, número de casos não desconhecidos e quartis. Testes: *H* de Kruskal-Wallis, mediana.

Considerações de Dados de Testes para Várias Amostras Independentes

Dados. Utilize variáveis numéricas que possam ser ordenadas.

Suposições. Utilize amostras aleatórias independentes. O teste *H* de Kruskal-Wallis requer que as amostras testadas sejam semelhantes na forma.

Para Obter Testes para Várias Amostras Independentes

1. Nos menus, escolha:

   Analisar > **Testes Não Paramétricos** > **Diálogos Anteriores** > **K Amostras Independentes ...**
2. Selecione uma ou mais variáveis numéricas.
3. Selecione uma variável de agrupamento e clique em **Definir Intervalo** para especificar valores de número inteiro mínimo e máximo para a variável de agrupamento.

Testes para vários tipos de teste de amostras independentes

Três testes estão disponíveis para determinar se várias amostras independentes vêm da mesma população. O teste *H* de Kruskal-Wallis, o teste de mediana e o Jonckheere-Terpstra, testam se as várias amostras independentes são da mesma população.

O **teste H de Kruskal-Wallis H**, uma extensão do teste *U* de Mann-Whitney, é o analógico não paramétrico da análise de variância para um fator que detecta as diferenças na localização da distribuição. O **teste de mediana**, que é um teste mais geral (mas não tão poderosa), detecta as diferenças de distribuição no local e formato. O teste *H* de Kruskal-Wallis e o teste de mediana assumem que não ha ordenação *a priori* das populações *k* das quais as amostras são colhidas.

Quando *houver* uma ordenação natural *a priori* (crescente ou decrescente) das populações *k*, o **teste Jonckheere-Terpstra** será mais poderoso. Por exemplo, as populações *k* podem representar *k* aumentando as temperaturas. A hipótese de que as temperaturas diferentes produzem a mesma distribuição de resposta é testada em relação à alternativa que, conforme a temperatura auenta, a magnitude da resposta aumenta também. Aqui, a hipótese alternativa é ordenada; portanto, Jonckheere-Terpstra é o teste mais apropriado para uso. O teste Jonckheere-Terpstra está disponível apenas se você tiver instalado o módulo complementar Testes Exatos.

Definir Intervalo para Testes de Diversas Amostras Independentes

Para definir o intervalo, insira valores de número inteiro para **Mínimo** e **Máximo** que correspondam às categorias mais baixa e mais alta da variável de agrupamento. Casos com valores fora dos limites são excluídos. Por exemplo, se especificar um valor mínimo de 1 e um valor máximo de 3, apenas os valores de número inteiro de 1 a 3 serão utilizados. O valor mínimo deve ser menor que o valor máximo, e ambos os valores devem ser especificados.

Opções de Testes para Várias Amostras Independentes

Estatísticas. É possível escolher uma ou ambas as estatísticas básicas.

- **Descritivo.** Exibe a média, o desvio padrão, o mínimo, o máximo e o número de casos não desconhecidos.
- **Quartis.** Exibe os valores correspondentes aos 25⌂, 50⌂ e 75⌂ percentis.

Valores omissos. Controla o tratamento de valores omissos.

- **Excluir casos de teste por teste.** Quando vários testes são especificados, cada teste é avaliado separadamente para valores omissos.
- **Excluir listwise dos casos.** Casos com valores omissos para qualquer variável são excluídos de todas as análises.

Recursos Adicionais do Comando NPAR TESTS (K Amostras Independentes)

A linguagem de sintaxe de comando também permite especificar um valor diferente da mediana observada para o teste de mediana (com o subcomando MEDIAN).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Testes para várias amostras relacionadas

O procedimento Testes para Várias Amostras Relacionadas compara as distribuições de duas ou mais variáveis.

Exemplo. O público associa diferentes quantias de prestígio a um médico, advogado, policial e professor? Dez pessoas são solicitadas a ranquear estas quatro ocupações em ordem de prestígio. O teste de Friedman indica que o público associa diferentes quantias de prestígio a essas quatro profissões.

Estatísticas. Média, desvio padrão, mínimo, máximo, número de casos não desconhecidos e quartis. Testes: Friedman, *W* de Kendall e *Q* de Cochran.

Considerações de Dados de Testes para Várias Amostras Relacionadas

Dados. Utilize variáveis numéricas que possam ser ordenadas.

Suposições. Testes não paramétricos não requerem suposições sobre a forma da distribuição subjacente. Utilize amostras aleatórias dependentes.

Para Obter Testes para Várias Amostras Dependentes

1. Nos menus, escolha:

   Analisar > **Testes Não Paramétricos** > **Diálogos Anteriores** > **K Amostras Relacionadas ...**
2. Selecionar duas ou mais variáveis de teste numéricas.

Testes para diversos tipos de teste de amostras relacionadas

Três testes estão disponíveis para comparar as distribuições de várias variáveis relacionadas.

O **teste** de **Friedman** é o equivalente não paramétrico de um design de medidas repetidas de uma amostra ou uma análise de variância de dois fatores com uma observação por célula. Friedman testa a hipótese nula de que *k* variáveis relacionadas vêm da mesma população. Para cada caso, as variáveis *k* são classificadas de 1 a *k*. A estatística de teste baseia-se nesses ranqueamentos.

W de Kendall é uma normalização da estatística de Friedman. *W* de Kendall é interpretável como o coeficiente de concordância, que é uma medida de acordo entre os avaliadores. Cada caso é um juiz ou avaliador, e cada variável é um item ou pessoa que está sendo avaliada. Para cada variável, a soma dos ranqueamentos é calculada. *W* de Kendall varia entre 0 (nenhuma concordância) e 1 (concordância total).

Q de Cochran é idêntico ao teste de Friedman, mas é aplicável quando todas as respostas são binárias. Este teste é uma extensão do teste de McNemar para a situação de amostra-*k*. *Q* de Cochran testa a hipótese de que diversas variáveis dicotômicas relacionadas têm a mesma média. As variáveis são medidas no mesmo indivíduo ou em indivíduos correspondidos.

Testes para Várias Estatísticas de Amostras Relacionadas

É possível escolher estatísticas.

- **Descritivo.** Exibe a média, o desvio padrão, o mínimo, o máximo e o número de casos não desconhecidos.
- **Quartis.** Exibe os valores correspondentes aos 25⌂, 50⌂ e 75⌂ percentis.

Recursos Adicionais do Comando NPAR TESTS (K Amostras Relacionadas)

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 28. Análise de Múltiplas Respostas

Análise de Múltiplas Respostas

Dois procedimentos estão disponíveis para analisar conjuntos de dicotomias múltiplas e de categorias múltiplas. O procedimento Frequências de Múltiplas Respostas exibe tabelas de frequência. O procedimento Crosstab de Múltiplas Respostas exibe tabulações cruzadas bidimensionais e tridimensionais. Antes de utilizar qualquer dos procedimentos, deve-se definir conjuntos de múltiplas respostas.

Exemplo. Este exemplo ilustra o uso de itens de múltiplas respostas em uma enquete de pesquisa de mercado. Os dados são fictícios e não devem ser interpretados como reais. Uma companhia aérea pode realizar uma pesquisa de opinião com passageiros que viajam em uma rota específica para avaliar as empresas concorrentes. Neste exemplo, a American Airlines deseja saber se seus passageiros também utilizam outras companhias aéreas para a rota Chicago - Nova York e se os horários e serviços são relativamente importantes na escolha de uma companhia aérea. Um comissário de bordo entrega para cada passageiro um breve questionário no momento do embarque. A primeira pergunta é: Circule todas as companhias aéreas com as quais você viajou pelo menos uma vez nos últimos seis meses nesta rota - American, United, TWA, USAir, Outra. Esta é uma pergunta de múltiplas respostas, já que o passageiro poderá circular mais de uma resposta. No entanto, essa pergunta não pode ser codificada diretamente porque uma variável pode ter somente um valor para cada caso. Deve-se utilizar diversas variáveis para mapear as respostas para cada pergunta. Existem duas maneiras de se fazer isso. Uma é definir uma variável que corresponda a cada uma das opções (por exemplo, American, United, TWA, USAir e Outra). Se o passageiro circular United, a variável *united* será designada a um código 1, caso contrário, 0. Este é um **método de múltiplas dicotomias** de mapeamento de variáveis. A outra maneira de mapear as respostas é o **método de múltiplas categorias**, em que você estima o número máximo de respostas possíveis para a pergunta e configura o mesmo número de variáveis, com códigos utilizados para especificar a companhia aérea utilizada. Ao analisar uma amostra de questionários, é possível descobrir que nenhum usuário utilizou mais de três companhias aéreas diferentes nesta rota nos últimos seis meses. Além disso, você descobre que, devido à desregulamentação de companhias aéreas, 10 outras empresas foram incluídas na categoria Outro. Utilizando o método de múltiplas respostas, você define três variáveis, cada uma codificada como 1 = *american*, 2 = *united*, 3 = *twa*, 4 = *usair*, 5 = *delta*, e assim por diante. Se um determinado passageiro circular American e TWA, a primeira variável terá um código de 1, a segunda terá um código de 3, e a terceira terá um código de valor omisso. Outro passageiro pode ter circulado American e inserido Delta. Assim, a primeira variável terá um código de 1, a segunda possuirá um código de 5, e a terceira um código de valor omisso. Se você utilizar o método de múltiplas dicotomias, por outro lado, acabará tendo 14 variáveis separadas. Embora qualquer método de mapeamento seja viável para essa pesquisa de opinião, o método que você escolher dependerá da distribuição das respostas.

Definir Conjuntos de Múltiplas Respostas

O procedimento Definir Conjuntos de Múltiplas Respostas agrupa variáveis elementares em conjuntos de diversas dicotomias e de diversas categorias para os quais é possível obter tabelas de frequência e tabulações cruzadas. É possível definir até 20 conjuntos de múltiplas respostas. Cada conjunto deve ter um nome exclusivo. Para remover um conjunto, destaque-o na lista de conjuntos de múltiplas respostas e clique em **Remover**. Para alterar um conjunto, destaque-o na lista, modifique quaisquer características de definições de conjunto e clique em **Alterar**.

É possível codificar suas variáveis elementares como dicotomias ou categorias. Para utilizar variáveis dicotômicas, selecione **Dicotomias** para criar um conjunto de múltiplas dicotomias. Insira um valor de número inteiro para o valor Contado. Cada variável que possuir pelo menos uma ocorrência do valor contado se torna uma categoria do conjunto de múltiplas dicotomias. Selecione **Categorias** para criar um

conjunto de categorias múltiplas com o mesmo intervalo de valores como variáveis de componente. Insira valores de número inteiro para os valores mínimo e máximo do intervalo para as categorias do conjunto de categorias múltiplas. O procedimento totaliza cada valor de número inteiro distinto no intervalo inclusivo em todas as variáveis de componente. Categorias vazias não são tabuladas.

Cada conjunto de múltiplas respostas deve ser designado a um nome exclusivo de até sete caracteres. O procedimento prefixa um sinal de dólar ($) ao nome que você designar. Não é possível usar os seguintes nomes reservados: *casenum*, *sysmis*, *jdate*, *date*, *time*, *length* e *width*. O nome do conjunto de múltiplas respostas existe apenas para uso em procedimentos de respostas múltiplas. Não é possível referenciar nomes de conjunto de múltiplas respostas em outros procedimentos. Opcionalmente, é possível inserir um rótulo de variável descritivo para o conjunto de múltiplas respostas. O rótulo pode ter até 40 caracteres de comprimento.

Para Definir Conjuntos de Múltiplas Respostas

1. Nos menus, escolha:
   Analisar > **Múltiplas Respostas** > **Defina Conjuntos de Variáveis...**
2. Selecione duas ou mais variáveis.
3. Se as variáveis forem codificadas como dicotomias, indique qual valor você deseja que seja contado. Se as variáveis forem codificadas como categorias, defina o intervalo das categorias.
4. Insira um nome exclusivo para cada conjunto de múltiplas respostas.
5. Clique em **Incluir** para incluir o conjunto de múltiplas respostas na lista de conjuntos definidos.

Frequências de múltiplas respostas

O procedimento Frequências de Múltiplas Respostas produz tabelas de frequência para conjuntos de múltiplas respostas. Primeiro deve-se definir um ou mais conjuntos de múltiplas respostas (consulte "Definir Conjuntos de Múltiplas Respostas").

Para conjuntos de múltiplas dicotomias, os nomes de categoria mostrados na saída são provenientes de rótulos de variáveis definidos para variáveis elementares no grupo. Se os rótulos de variáveis não forem definidos, nomes de variáveis serão utilizados como rótulos. Para conjuntos de categorias múltiplas, os rótulos de categoria são provenientes de rótulos de valor da primeira variável no grupo. Se categorias omissas para a primeira variável estiverem presentes para outras variáveis no grupo, defina um rótulo de valor para as categorias omissas.

Valores omissos. Casos com valores omissos são excluídos basicamente tabela por tabela. Como alternativa, é possível escolher um ou ambos os seguintes itens:

- **Excluir listwise dos casos dentro das dicotomias.** Exclui casos com valores omissos para qualquer variável da tabulação do conjunto de múltiplas dicotomias. Isso se aplica apenas a conjuntos de múltiplas respostas definidos como conjuntos de dicotomias. Por padrão, um caso será considerado omisso para um conjunto de múltiplas dicotomias se nenhuma de suas variáveis de componente contiver o valor contado. Casos com valores omissos para algumas (mas não todas) variáveis são incluídos nas tabulações do grupo se pelo menos uma variável contiver o valor contado.
- **Excluir listwise dos casos dentro das categorias.** Exclui casos com valores omissos para qualquer variável da tabulação do conjunto de categorias múltiplas. Isso se aplica apenas a conjuntos de múltiplas respostas definidos como conjuntos de categoria. Por padrão, um caso será considerado omisso para um conjunto de categorias múltiplas apenas se nenhum de seus componentes possuir valores válidos dentro do intervalo definido.

Exemplo. Cada variável criada a partir de uma pergunta da pesquisa de opinião é uma variável elementar. Para analisar um item de múltiplas respostas, deve-se combinar as variáveis em um dos dois tipos de conjuntos de múltiplas respostas: um conjunto de múltiplas dicotomias ou um conjunto de categorias múltiplas. Por exemplo, se uma pesquisa de opinião de uma companhia aérea perguntar com quais das três companhias (American, United e TWA) você viajou nos últimos seis meses e você utilizou

as variáveis dicotômicas e definiu um **conjunto de múltiplas dicotomias**, cada uma das três variáveis no conjunto se tornará uma categoria da variável de grupo. As contagens e porcentagens para as três companhias aéreas são exibidas em uma tabela de frequência. Se você descobrir que nenhum respondente mencionou mais de duas companhias aéreas, será possível criar duas variáveis, cada uma com três códigos, um para cada companhia aérea. Se você definir um **conjunto de categorias múltiplas**, os valores serão tabulados ao incluir os mesmos códigos nas variáveis elementares juntas. O conjunto de valores resultante é o mesmo que aqueles para cada uma das variáveis elementares. Por exemplo, 30 respostas para United são a soma das cinco respostas de United para a companhia aérea 1 e as 25 respostas de United para a companhia aérea 2. As contagens e porcentagens para as três companhias aéreas são exibidas em uma tabela de frequência.

Estatísticas. Tabelas de frequência exibindo contagens, as porcentagens de respostas, porcentagens de casos, número de casos válidos e o número de casos omissos.

Considerações de Dados de Frequências de Múltiplas Respostas

Dados. Utilize conjuntos de múltiplas respostas.

Suposições. As contagens e porcentagens fornecem uma descrição útil dos dados a partir de qualquer distribuição.

Procedimentos relacionados. O procedimento Definir Conjuntos de Múltiplas Respostas permite definir conjuntos de múltiplas respostas.

Para Obter Frequências de Múltiplas Respostas

1. Nos menus, escolha:

   Analisar > **Múltiplas Respostas** > **Frequências...**
2. Selecione um ou mais conjuntos de múltiplas respostas.

Tabulações cruzadas de múltiplas respostas

O procedimento Crosstabs de Múltiplas Respostas efetua tabulações cruzadas de conjuntos de múltiplas respostas definidos, de variáveis elementares, ou uma combinação deles. Também é possível obter porcentagens de célula com base em casos ou respostas, modificar a manipulação de valores omissos ou obter tabulações cruzadas pairwise. Primeiro deve-se definir um ou mais conjuntos de múltiplas respostas (consulte "Para Definir Conjuntos de Múltiplas Respostas").

Para conjuntos de múltiplas dicotomias, os nomes de categoria mostrados na saída são provenientes de rótulos de variáveis definidos para variáveis elementares no grupo. Se os rótulos de variáveis não forem definidos, nomes de variáveis serão utilizados como rótulos. Para conjuntos de categorias múltiplas, os rótulos de categoria são provenientes de rótulos de valor da primeira variável no grupo. Se categorias omissas para a primeira variável estiverem presentes para outras variáveis no grupo, defina um rótulo de valor para as categorias omissas. O procedimento exibe rótulos de categoria para colunas em três linhas, com até oito caracteres por linha. Para evitar divisão de palavras, é possível reverter os itens de linha e de coluna ou redefinir rótulos.

Exemplo. Os conjuntos de múltiplas dicotomias e de múltiplas categorias podem efetuar tabulação cruzada com outras variáveis neste procedimento. Uma pesquisa de opinião para passageiros de companhias aéreas pede o seguinte: Circule todas as companhias aéreas com as quais você viajou pelo menos uma vez nos últimos seis meses (American, United, TWA). O que é mais importante na escolha de um voo - horário ou serviço? Selecione apenas um. Após inserir os dados como múltiplas dicotomias ou categorias e combiná-las em um conjunto, é possível efetuar uma tabulação cruzada das opções da companhia aérea com a pergunta que envolve serviço ou horário.

Estatísticas. Tabulação cruzada com célula, linha, coluna e contagens totais, e com célula, linha, coluna e porcentagens totais. As porcentagens de célula podem ser baseadas em casos ou respostas.

Considerações de Dados para Crosstabs de Múltiplas Respostas

Dados. Utilize conjuntos de múltiplas respostas ou variáveis categóricas numéricas.

Suposições. As contagens e porcentagens fornecem uma descrição útil dos dados a partir de qualquer distribuição.

Procedimentos relacionados. O procedimento Definir Conjuntos de Múltiplas Respostas permite definir conjuntos de múltiplas respostas.

Para Obter Crosstabs de Múltiplas Respostas

1. Nos menus, escolha:
   Analisar > **Múltiplas Respostas** > **Crosstabs...**
2. Selecione uma ou mais variáveis numéricas ou conjuntos de múltiplas respostas para cada dimensão da tabulação cruzada.
3. Defina o intervalo de cada variável elementar.

Opcionalmente, é possível obter uma tabulação cruzada bidirecional para cada categoria de uma variável de controle ou conjunto de múltiplas respostas. Selecione um ou mais itens para a lista de Camada(s).

Intervalos de Definição de Crosstabs de Múltiplas Respostas

Intervalos de valor devem ser definidos para qualquer variável elementar na tabulação cruzada. Insira os valores de categoria mínimo e máximo de número inteiro que deseja tabular. Categorias fora do intervalo são excluídas da análise. Valores dentro do intervalo inclusivo são assumidos como números inteiros (não inteiros são truncados).

Opções de Crosstabs de Múltiplas Respostas

Porcentagens de Célula. As contagens de célula são sempre exibidas. É possível optar por exibir porcentagens de linha, porcentagens de coluna e porcentagens de tabela de dois fatores (total).

Porcentagens Baseadas em. É possível basear porcentagens de célula em casos (ou respondentes). Isso não estará disponível se selecionar correspondência de variáveis em conjuntos de categorias múltiplas. Também é possível basear as porcentagens de célula em respostas. Para conjuntos de múltiplas dicotomias, o número de respostas é igual ao número de valores contabilizados nos casos. Para conjuntos de categorias múltiplas, o número de respostas é o número de valores no intervalo definido.

Valores omissos. É possível escolher uma ou ambas as seguintes opções:

- **Excluir listwise dos casos dentro das dicotomias.** Exclui casos com valores omissos para qualquer variável da tabulação do conjunto de múltiplas dicotomias. Isso se aplica apenas a conjuntos de múltiplas respostas definidos como conjuntos de dicotomias. Por padrão, um caso será considerado omisso para um conjunto de múltiplas dicotomias se nenhuma de suas variáveis de componente contiver o valor contado. Casos com valores omissos para algumas, mas não todas as variáveis são incluídas nas tabulações do grupo se pelo menos uma variável contiver o valor contado.
- **Excluir listwise dos casos dentro das categorias.** Exclui casos com valores omissos para qualquer variável da tabulação do conjunto de categorias múltiplas. Isso se aplica apenas a conjuntos de múltiplas respostas definidos como conjuntos de categoria. Por padrão, um caso será considerado omisso para um conjunto de categorias múltiplas apenas se nenhum de seus componentes possuir valores válidos dentro do intervalo definido.

Por padrão, ao efetuar tabulação cruzada de dois conjuntos de categorias múltiplas, o procedimento irá tabular cada variável no primeiro grupo com cada variável no segundo grupo e somará as contagens de cada célula; portanto, algumas respostas podem aparecer mais de uma vez em uma tabela. É possível escolher a seguinte opção:

Corresponder variáveis em conjuntos de respostas. Emparelha a primeira variável no primeiro grupo com a primeira variável no segundo grupos, e assim por diante. Se essa opção for selecionada, o procedimento baseará as porcentagens da célula nas respostas e não nos respondentes. O emparelhamento não está disponível para conjuntos de múltiplas dicotomias ou variáveis elementares.

Recursos Adicionais do Comando MULT RESPONSE

O idioma da sintaxe de comando também permite:

- Obter tabelas de tabulação cruzada com até cinco dimensões (com o subcomando `BY`).
- Alterar as opções de formatação de saída, incluindo a supressão de rótulos de valor (com o subcomando `FORMAT`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 29. Resultados do Relatório

Resultados do Relatório

Listagens de casos e estatísticas descritivas são ferramentas básicas para estudar e apresentar dados. É possível obter listagens de caso com o Editor de Dados ou com o procedimento Sumarizar, contagens de frequência e estatísticas descritivas com o procedimento Frequências e as estatísticas de subpopulação com o procedimento Médias. Cada um desses utiliza um formato projetado para tornar as informações claras. Se desejar exibir as informações em um formato diferente, as Sumarizações de Relatório em Linhas e Sumarizações de Relatório em Colunas fornecem controle sobre a apresentação de dados.

Sumarizações de Relatórios em Linhas

As Sumarizações de Relatórios em Linhas produz relatórios em que estatísticas básicas diferentes são dispostas em linhas. Listagens de casos também estão disponíveis, com ou sem estatísticas básicas.

Exemplo. Uma empresa com uma rede de lojas varejistas mantém registros de informações de funcionários, incluindo salário, tempo de serviço e a loja e divisão em que cada funcionário trabalha. É possível gerar um relatório que forneça informações de funcionário individual (listando), discriminado por loja e divisão (variáveis de quebra), com estatísticas básicas (por exemplo, média salarial) para cada loja, divisão e a divisão de cada loja.

Colunas de Dados. Lista as variáveis de relatório para as quais você deseja listagens de caso ou estatísticas básicas e controla o formato de exibição de colunas de dados.

Colunas de Quebra. Lista as variáveis de quebra opcionais que dividem o relatório em grupos e controla as estatísticas básicas e os formatos de exibição das colunas de quebra. Para diversas variáveis de quebra, haverá um grupo separado para cada categoria de cada variável de quebra dentro de categorias da variável de quebra anterior na lista. As variáveis de quebra devem ser variáveis categóricas discretas que dividem casos em um número limitado de categorias significativas. Valores individuais de cada variável de quebra aparecem, ordenados em uma coluna separada à esquerda de todas as colunas de dados.

Relatório. Controla as características gerais do relatório, incluindo estatísticas básicas gerais, exibição de valores omissos, numeração de página e títulos.

Exibir casos. Exibe os valores reais (ou rótulos de valor) as variáveis de coluna de dados para cada caso. Isso produz um relatório de listagem, que pode ser muito mais longo que um relatório sumarização.

Visualizar. Exibe apenas a primeira página do relatório. Esta opção é útil para visualizar o formato de seu relatório sem processar o relatório inteiro.

Dados já estão ordenados. Para relatórios com variáveis de quebra, o arquivo de dados deve ser ordenado pelos valores de variável de quebra antes de gerar o relatório. Se o seu arquivo de dados já estiver ordenado pelos valores de variáveis de quebra, será possível economizar tempo de processamento ao selecionar essa opção. Essa opção é útil principalmente após a execução de um relatório de visualização.

Para Obter um Relatório Sumarização: Sumarizações em Linhas

1. Nos menus, escolha:

   Analisar > **Relatórios** > **Sumarizações de Relatório em Linhas...**
2. Selecione uma ou mais variáveis para Colunas de Dados. Uma coluna no relatório é gerada para cada variável selecionada.

3. Para relatórios ordenados e exibidos por subgrupos, selecione mais variáveis para Colunas de Quebra.
4. Para relatórios com estatísticas básicas para subgrupos definidos pelas variáveis de quebra, selecione a variável de quebra na lista de Variáveis de Colunas de Quebra e clique em **Sumarização** no grupo de Colunas de Quebra para especificar uma ou mais medidas de sumarização.
5. Para relatórios com estatísticas básicas gerais, clique em **Sumarização** para especificar uma ou mais medidas de sumarização.

Formato de Quebra de Coluna de Dados do Relatório

As caixas de diálogo Formato controlam os títulos da coluna, a largura da coluna, o alinhamento de texto e a exibição de valores de dados ou de rótulos de valor. O Formato de Coluna de Dados controla o formato de colunas de dados no lado direito da página do relatório. O Formato de Quebra controla o formato das colunas de quebra no lado esquerdo.

Título da Coluna. Para a variável selecionada, controla o título da coluna. Títulos longos são agrupados automaticamente dentro da coluna. Use a tecla Enter para inserir manualmente quebras de linha no local desejado para quebra de títulos.

Posição do Valor dentro da Coluna. Para a variável selecionada, controla o alinhamento de valores de dados ou de rótulos de valor dentro da coluna. O alinhamento de valores ou de rótulos não afeta o alinhamento de títulos de coluna. É possível indentar o conteúdo da coluna por um número especificado de caracteres ou centralizar o conteúdo.

Conteúdo da Coluna. Para a variável selecionada, controla a exibição de valores de dados ou de rótulos de valor definido. Os valores de dados são sempre exibidos para quaisquer valores que não possuírem rótulos de valor definido. (Não disponível para colunas de dados nos relatórios de sumarização da coluna).

Linhas de Sumarização de Relatório para/Linhas de Sumarização Final

As duas caixas de diálogo Linhas de Sumarização controlam a exibição das estatísticas básicas para grupos de quebra e para o relatório inteiro. Linhas de Sumarização controla estatísticas de subgrupo para cada categoria definida por uma ou mais variáveis de quebra. Linhas de Sumarização Final controla as estatísticas globais, exibidas no término do relatório.

As estatísticas básicas disponíveis são soma, média, mínimo, máximo, número de casos, porcentagem de casos acima ou abaixo de um valor especificado, porcentagem de casos dentro de um intervalo especificado de valores, desvio padrão, curtose, variância e assimetria.

Opções de Quebra de Relatório

Opções de Quebra controlam o espaçamento e a paginação de informações sobre categoria de quebra.

Controle da Página. Controla o espaçamento e a paginação para categorias da variável de quebra selecionada. É possível especificar um número de linhas em branco entre as categorias de quebra ou iniciar cada categoria de quebra em uma nova página.

Linhas em Branco antes de Sumarizações. Controla o número de linhas em branco entre os rótulos de categoria de quebra ou estatísticas básicas e de dados. Isto é útil principalmente para relatórios combinados que incluem tanto listagens de caso individuais quanto estatísticas básicas para as categorias de quebra; nesses relatórios, é possível inserir espaço entre as listagens de caso e as estatísticas básicas.

Opções de relatório

Opções de Relatório controla o tratamento e a exibição de valores omissos e a numeração de página de relatório.

Excluir casos com listwise de valores omissos. Elimina (do relatório) qualquer caso com valores omissos para qualquer uma das variáveis de relatório.

Valores Omissos Aparecem como. Permite especificar o símbolo que representa valores omissos no arquivo de dados. O símbolo pode ser apenas um caractere e é utilizado para representar valores *omissos do sistema* e *omissos do usuário*.

Numerar Páginas a partir de. Permite especificar um número de página para a primeira página do relatório.

Layout de relatórios

O Layout do Relatório controla a largura e o comprimento de cada página de relatório, o posicionamento do relatório na página e a inserção de linhas e rótulos em branco.

Formato da Página. Controla as margens da página, expressas em linhas (superior e inferior) e caracteres (esquerda e direita), e o alinhamento de relatórios dentro das margens.

Títulos e Rodapés da Página. Controla o número de linhas que separam títulos e rodapés da página do corpo do relatório.

Colunas de Quebra. Controla a exibição de colunas de quebra. Se diversas variáveis de quebra forem especificadas, elas poderão ser separadas em colunas ou na primeira coluna. Colocar todas as variáveis de quebra na primeira coluna produz um relatório mais estreito.

Títulos da Coluna. Controla a exibição de títulos de coluna, incluindo título sublinhado, espaço entre os títulos e o corpo do relatório, e o alinhamento vertical dos títulos da coluna.

Rótulos de Linhas e de Quebra de Colunas de Dados. Controla o posicionamento das informações da coluna de dados (valores de dados e/ou estatísticas básicas) com relação aos rótulos de quebra no início de cada categoria de quebra. A primeira linha de informações de coluna de dados pode iniciar na mesma linha que o rótulo de categoria de quebra ou em um número especificado de linhas após o rótulo de categoria de quebra. (Não disponível para colunas de relatório sumarização).

Títulos de Relatório

Títulos de Relatório controlam o conteúdo e o posicionamento dos títulos e rodapés do relatório. É possível especificar até 10 linhas de títulos de página e até 10 linhas de rodapés de página, com componentes justificados à esquerda, centralizados e justificados à direita em cada linha.

Se inserir variáveis nos títulos ou rodapés, o rótulo de valor ou o valor da variável atual será exibido no título ou no rodapé. Nos títulos, o rótulo de valor correspondente ao valor da variável no início da página é exibido. Nos rodapés, o rótulo de valor correspondente ao valor da variável no término da página é exibido. Se não houver nenhum rótulo de valor, o valor real será exibido.

Variáveis Especiais. As variáveis especiais *DATE* e *PAGE* permitem inserir a data ou o número da página atual em qualquer linha de um cabeçalho ou rodapé de relatório. Se o arquivo de dados contiver variáveis denominadas *DATE* ou *PAGE*, não será possível utilizar essas variáveis em títulos ou rodapés do relatório.

Sumarizações de Relatórios em Colunas

As Sumarizações de Relatórios em Colunas produz relatórios sumarização em que estatísticas básicas diferentes aparecem em colunas separadas.

Exemplo. Uma empresa com uma rede de lojas varejistas mantém registros de informações de funcionários, incluindo salário, tempo de serviço e a divisão em que cada funcionário trabalha. É possível gerar um relatório que fornece estatísticas de salário sumarizadas (por exemplo, média, mínimo e máximo) para cada divisão.

Colunas de Dados. Lista as variáveis de relatório para as quais você deseja estatísticas básicas e controla o formato de exibição e as estatísticas básicas exibidas para cada variável.

Colunas de Quebra. Lista as variáveis de quebra opcionais que dividem o relatório em grupos e controla os formatos de exibição das colunas de quebra. Para diversas variáveis de quebra, haverá um grupo separado para cada categoria de cada variável de quebra dentro de categorias da variável de quebra anterior na lista. As variáveis de quebra devem ser variáveis categóricas discretas que dividem casos em um número limitado de categorias significativas.

Relatório. Controla as características gerais do relatório, incluindo exibição de valores omissos, numeração de página e títulos.

Visualizar. Exibe apenas a primeira página do relatório. Esta opção é útil para visualizar o formato de seu relatório sem processar o relatório inteiro.

Dados já estão ordenados. Para relatórios com variáveis de quebra, o arquivo de dados deve ser ordenado pelos valores de variável de quebra antes de gerar o relatório. Se o seu arquivo de dados já estiver ordenado pelos valores de variáveis de quebra, será possível economizar tempo de processamento ao selecionar essa opção. Essa opção é útil principalmente após a execução de um relatório de visualização.

Para Obter um Relatório Sumarização: Sumarizações em Colunas

1. Nos menus, escolha:

   Analisar > **Relatórios** > **Sumarizações de Relatório em Colunas...**
2. Selecione uma ou mais variáveis para Colunas de Dados. Uma coluna no relatório é gerada para cada variável selecionada.
3. Para alterar a medida de sumarização de uma variável, selecione a variável na lista Variáveis de Coluna de Dados e clique em **Sumarização**.
4. Para obter mais de uma medida de sumarização para uma variável, selecione a variável na lista de origem e mova-a para a lista Variáveis de Coluna de Dados diversas vezes, uma para cada medida de sumarização desejada.
5. Para exibir uma coluna contendo a soma, média, razão ou outra função de colunas existentes, clique em **Inserir Total**. Isso coloca uma variável denominada *total* na lista Colunas de Dados.
6. Para relatórios ordenados e exibidos por subgrupos, selecione mais variáveis para Colunas de Quebra.

Função de Sumarização de Colunas de Dados

Linhas de Sumarização controla a estatística de sumarização exibida para variável da coluna de dados selecionada.

As estatísticas básicas disponíveis são soma, média, mínimo, máximo, número de casos, porcentagem de casos acima ou abaixo de um valor especificado, porcentagem de casos dentro de um intervalo especificado de valores, desvio padrão, variância, curtose e assimetria.

Sumarização de Colunas de Dados para Total de Colunas

Coluna de Sumarização controla as estatísticas básicas totais que sumarizam duas ou mais colunas de dados.

As estatísticas básicas totais disponíveis são soma de colunas, média de colunas, mínimo, máximo, diferença entre valores em duas colunas, quociente de valores em uma coluna dividido por valores em outra coluna, e o produto dos valores de colunas multiplicados juntos.

Soma de colunas. A coluna *total* é a soma das colunas na lista Colunas de Sumarização.

Média de colunas. A coluna *total* é a média das colunas na lista Colunas de Sumarização.

Mínimo de colunas. A coluna *total* é o mínimo das colunas na lista de Colunas de Sumarização.

Máximo de colunas. A coluna *total* é o máximo das colunas na lista Colunas de Sumarização.

1ª coluna - 2ª coluna. A coluna *total* é a diferença das colunas na lista de Colunas de Sumarização. A lista de Coluna de Sumarização deve conter exatamente duas colunas.

1ª coluna / 2ª coluna. A coluna *total* é o quociente das colunas na lista Colunas de Sumarização. A lista de Coluna de Sumarização deve conter exatamente duas colunas.

% 1ª coluna / 2ª coluna. A coluna *total* é a porcentagem da primeira coluna da segunda coluna na lista Coluna de Sumarização. A lista de Coluna de Sumarização deve conter exatamente duas colunas.

Produto de colunas. A coluna *total* é o produto das colunas na lista de Coluna de Sumarização.

Formato da Coluna do Relatório

As opções de formatação de dados e coluna de quebra para Sumarizações de Relatórios em Colunas são as mesmas que aquelas descritas para Sumarizações de Relatórios em Linhas.

Sumarizações de Relatórios em Opções de Quebra de Colunas

As Opções de Quebra controlam a exibição de subtotal, espaçamento e paginação de categorias de quebra.

Subtotal. Controla a exibição de subtotais para as categorias de quebra.

Controle da Página. Controla o espaçamento e a paginação para categorias da variável de quebra selecionada. É possível especificar um número de linhas em branco entre as categorias de quebra ou iniciar cada categoria de quebra em uma nova página.

Linhas em Branco antes do subtotal. Controla o número de linhas em branco entre os dados e subtotais da categoria de quebra.

Sumarizações de Relatórios em Opções de Colunas

Opções controla a exibição de totais gerais, a exibição de valores omissos e a paginação em relatórios sumarização de coluna.

Total Geral. Exibe e rotula um total geral para cada coluna; exibido na parte inferior da coluna.

Valores omissos. É possível excluir valores omissos do relatório ou selecionar um único caractere para indicar valores omissos no relatório.

Layout de Relatório para Sumarizações nas colunas

As opções de layout de relatório para Sumarizações de Relatórios em Colunas são as mesmas que aquelas descritas para Sumarizações de Relatórios em Linhas.

Recursos Adicionais do Comando REPORT

O idioma da sintaxe de comando também permite:

- Exibir funções de sumarização diferentes nas colunas de uma linha de sumarização única.
- Inserir linhas de sumarização em colunas de dados para variáveis diferentes de variável de coluna de dados ou para várias combinações (funções compostas) das funções de sumarização.
- Utilizar Mediana, Modo, Frequência e Porcentagem como funções de sumarização.
- Controlar com mais precisão o formato de exibição de estatísticas de sumarização.
- Inserir linhas em branco em vários pontos em relatórios.
- Inserir linhas em branco após cada *n*-ésimo caso nos relatórios de listagem.

Devido à complexidade da sintaxe `REPORT`, talvez você ache útil, ao construir um novo relatório com a sintaxe, aproximar o relatório gerado das caixas de diálogo, copiar e colar a sintaxe correspondente e refinar essa sintaxe para produzir o relatório exato desejado.

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 30. Análise de confiabilidade

A análise de confiabilidade permite estudar as propriedades das escalas de medida e os itens que compõem as escalas. O procedimento Análise de Confiabilidade calcula um número de medidas normalmente utilizadas de confiabilidade da escala e também fornece informações sobre os relacionamentos entre itens individuais na escala. Os coeficientes de correlação intraclasse podem ser utilizados para calcular estimativas de confiabilidade entre avaliadores.

Exemplo. O meu questionário mede a satisfação do cliente de maneira útil? Utilizando a análise de confiabilidade, é possível determinar até que ponto os itens em seu questionário estão relacionados uns aos outros, obter um índice geral da repetição e consistência interna da escala como um todo e identificar itens de problemas que devem ser excluídos da escala.

Estatísticas. Descritivas para cada variável e para a escala, estatísticas básicas em itens, correlações e covariâncias entre itens, estimativas de confiabilidade, tabela de ANOVA, coeficientes de correlação intraclasse, T^2 de Hotelling e teste de aditividade de Tukey.

Modelos. Os modelos de confiabilidade a seguir estão disponíveis:

- **Alfa (Cronbach).** Este modelo é um modelo de consistência interna, com base na correlação média entre itens.
- **Duas Metades.** Esse modelo divide a escala em duas partes e examina a correlação entre as partes.
- **Guttman.** Esse modelo calcula limites inferiores de Guttman para confiabilidade real.
- **Paralelo.** Esse modelo supõe que todos os itens possuem variâncias e variâncias de erro iguais entre as replicações.
- **Paralelo-série.** Esse modelo faz suposições do modelo Paralelo e também supõe médias iguais entre os itens.

Considerações de Dados de Análise de Confiabilidade

Dados. Os dados podem ser dicotômicos, ordinais ou intervalo, porém devem ser codificados numericamente.

Suposições. As observações devem ser independentes, e os erros deverão estar não correlacionados entre os itens. Cada par de itens deve ter uma distribuição normal bivariada. As escalas devem ser aditivas, para que cada item esteja linearmente relacionado ao escore total.

Procedimentos relacionados. Se desejar explorar a dimensionalidade de seus itens de escala (para ver se mais de uma construção é necessária para considerar o padrão de escores de item), use a análise fatorial ou o ajuste de escala multidimensional. Para identificar grupos homogêneos de variáveis, utilize análise de cluster hierárquica para variáveis de cluster.

Para Obter uma Análise de Confiabilidade

1. Nos menus, escolha:

   Analisar > **Escala** > **Análise de Confiabilidade...**
2. Selecione duas ou mais variáveis como componentes em potencial de uma escala aditiva.
3. Escolha um modelo na lista suspensa de Modelo.

Estatística de Análise de Confiabilidade

É possível selecionar várias estatísticas que descrevem a escala e os itens. As estatísticas que são relatadas por padrão incluem o número de casos, o número de itens e estimativas de confiabilidade conforme a seguir:

- **Modelos alfa.** Coeficiente alfa; para dados dicotômicos, isto é equivalente ao coeficiente de Kuder-Richardson 20 (KR20).
- **Modelo split-half.** Correlação entre formas, confiabilidade split-half de Guttman, confiabilidade de Spearman-Brown (comprimento igual e desigual) e o coeficiente alfa para cada metade.
- **Modelos de Guttman.** Coeficientes de confiabilidade lambda 1 a lambda 6.
- **Modelos Paralelo e Paralelo-Série.** Teste para Qualidade do ajuste do modelo, estimativas de variância do erro, variância comum e variância verdadeira, correlação entre itens comum estimada, confiabilidade estimada e estimativa imparcial da confiabilidade.

Descritivos para. Produz estatísticas descritivas para escalas ou itens entre os casos.

- **Item.** Produz estatísticas descritivas para os itens entre os casos.
- **Escala.** Produz estatísticas descritivas para as escalas.
- **Escalar se item excluído.** Exibe estatísticas básicas comparando cada item com a escala que é composta dos outros itens. As estatísticas incluem a média e a variância de escala, caso o item tivesse que ser excluído da escala, a correlação entre o item e a escala que é composta de outros itens e o alfa de Cronbach, caso o item tivesse que ser excluído da escala.

Sumarizações. Fornece estatísticas descritivas das distribuições do item em todos os itens na escala.

- *Médias*. Estatística de sumarização para médias de item. São exibidas as médias de item maiores, menores e médias, o intervalo e a variância das médias de item e a razão das médias de item, das maiores às menores.
- *Variâncias*. Estatísticas de sumarização para variâncias de item. As variâncias de item maiores, menores e médias, o intervalo e a variância de variâncias de item e a razão das variâncias de item maiores com as menores são exibidos.
- *Covariâncias*. Estatísticas de sumarização para covariâncias entre itens. As covariâncias entre itens maiores, menores e médias, o intervalo e a variância de covariâncias entre itens e a razão das covariâncias entre itens maiores com as menores são exibidos.
- *Correlações*. Estatísticas de sumarização para correlações entre itens. As correlações entre itens maiores, menores e médias, o intervalo e a variância de correlações entre itens e a razão das correlações entre itens maiores com as menores são exibidos.

Entre Itens. Produz matrizes de correlações ou de covariâncias entre os itens.

Tabela ANOVA. Produz testes de médias iguais.

- *Teste de F*. Exibe uma tabela de análise de variância de medidas repetidas.
- *Qui-quadrado de Friedman*. Exibe o qui-quadrado de Friedman e o coeficiente de concordância de Kendall. Essa opção é apropriada para os dados que estiverem na forma de ranqueamentos. O teste qui-quadrado substitui o teste F usual na tabela ANOVA.
- *Qui-quadrado de Cochran*. Exibe o Q de Cochran. Essa opção é apropriada para dados que forem dicotômicos. A estatística Q substitui a estatística F usual na tabela ANOVA.

T Quadrado de Hotelling. Produz um teste multivariado da hipótese nula de que todos os itens na escala têm a mesma média.

Teste de aditividade de Tukey Produz um teste da suposição de que não há nenhuma interação multiplicativa entre os itens.

Coeficiente de correlação intraclasse. Produz medidas de consistência ou acordo de valores dentro dos casos.

- **Modelo.** Selecione o modelo para calcular o coeficiente de correlação intraclasse. Os modelos disponíveis são Bilateral Combinado, Bilateral Aleatório e Unilateral Aleatório. Selecione **Bilateral Combinado** quando os efeitos de pessoas são aleatórios e os efeitos de item são fixos, selecione **Bilateral Aleatório** quando os efeitos de pessoas e os efeitos de item são aleatórios, ou selecione **Unilateral Aleatório** quando os efeitos de pessoas são aleatórios.
- **Tipo.** Selecione o tipo de índice. Os tipos disponíveis são Consistência e Acordo Absoluto.
- **Intervalo de confiança.** Especifique o nível do intervalo de confiança. O padrão é 95%.
- **Valor de teste.** Especifique o valor hipotético do coeficiente para o teste de hipótese. Este valor é o valor com o qual o valor observado é comparado. O valor padrão é 0.

Recursos Adicionais do Comando RELIABILITY

O idioma da sintaxe de comando também permite:

- Ler e analisar uma matriz de correlações.
- Gravar uma matriz de correlações para análise posterior.
- Especificar divisões diferentes de metades iguais para o método das duas metades.

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 31. Escala multidimensional

O ajuste de escala multidimensional tenta localizar a estrutura em um conjunto de medidas de distância entre os objetos ou casos. Essa tarefa é realizada ao designar observações para localizações específicas em um espaço conceitual (geralmente de duas ou três dimensões), de forma que as distâncias entre os pontos no espaço correspondem às dissimilaridades especificadas o mais próximo possível. Em muitos casos, as dimensões deste espaço conceitual podem ser interpretadas e utilizadas para entender ainda mais os seus dados.

Se você tiver variáveis medidas de modo objetivo, será possível utilizar o ajuste de escala multidimensional como uma técnica de redução de dados (o procedimento de Ajuste de Escala Multidimensional calculará distâncias de dados multivariados, se necessário). O ajuste de escala multidimensional também pode ser aplicado às classificações subjetivas da dissimilaridade entre os objetos ou conceitos. Além disso, o procedimento Ajuste de Escala Multidimensional pode manipular dados de dissimilaridade a partir de diversas origens, assim como pode ocorrer com diversos avaliadores ou respondentes de questionários.

Exemplo. Como as pessoas percebem relacionamentos entre diferentes automóveis? Se você tiver dados de respondentes indicando classificações de similaridade entre diferentes marcas e modelo de automóveis, o ajuste de escala multidimensional pode ser utilizado para identificar as dimensões que descrevem as percepções dos consumidores. Você pode achar, por exemplo, que o preço e o tamanho de um veículo definem um espaço bidimensional, que justificam as similaridades relatadas pelos seus respondentes.

Estatísticas. Para cada modelo: matriz de dados, matriz de dados idealmente escalada, estresse de S (Young), estresse (Kruskal), RSQ, coordenadas de estímulo, estresse médio e RSQ para cada estímulo (modelos RMDS). Para modelos de diferença individuais (INDSCAL): ponderações de assuntos e de índice de anomalia para cada assunto. Para cada matriz em modelos de ajuste de escala multidimensional replicados: estresse e RSQ para cada estímulo. Gráficos: coordenadas de estímulo (duas ou três dimensões), gráfico de dispersão de disparidades versus distâncias.

Considerações de Dados para Ajuste de Escala Multidimensional

Dados. Se seus dados forem dados de dissimilaridade, todas as dissimilaridades deverão ser quantitativas e medidas na mesma métrica. Se seus dados forem dados multivariados, as variáveis poderão ser quantitativas, binárias ou dados de contagem. O ajuste de escala de variáveis é uma consideração importante - as diferenças no ajuste de escala podem afetar sua solução. Se as suas variáveis possuírem diferenças grandes no ajuste de escala (por exemplo, uma variável é medida em dólares e a outra variável é medida em anos), considere padronizá-las (esse processo pode ser feito automaticamente pelo procedimento Ajuste de Escala Multidimensional).

Suposições. O procedimento Ajuste de Escala Multidimensional é relativamente livre de suposições de distribuição. Assegure-se de selecionar o nível de medição apropriado (ordinal, intervalo ou proporção) na caixa de diálogo Opções de Ajuste de Escala Multidimensional para que os resultados sejam calculados corretamente.

Procedimentos relacionados. Se o seu objetivo é redução de dados, um método alternativo a ser considerado é a análise fatorial, principalmente se suas variáveis forem quantitativas. Se desejar identificar grupos de casos semelhantes, considere complementar o ajuste de escala multidimensional com uma análise de cluster hierárquico ou por *k*-médias.

Para Obter uma Análise de Ajuste de Escala Multidimensional

1. Nos menus, escolha:

Analisar > **Escala** > **Ajuste de Escala Multidimensional...**

2. Selecione pelo menos quatro variáveis numéricas para análise.
3. No grupo Distâncias, selecione **Dados são distâncias** ou **Criar distâncias a partir de dados**.
4. Se selecionar **Criar distâncias a partir de dados**, também será possível selecionar uma variável de agrupamento para matrizes individuais. A variável de agrupamento pode ser numérica ou de sequência de caracteres.

Opcionalmente, também é possível:

- Especificar a forma da matriz de distância quando os dados são distâncias.
- Especificar a medida de distância a ser utilizada ao criar as distâncias a partir dos dados.

Forma de Dados para Ajuste de Escala Multidimensional

Se o seu conjunto de dados ativo representar distâncias entre um conjunto de objetos ou representar distâncias entre dois conjuntos de objetos, especifique a forma de sua matriz de dados para obter os resultados corretos.

Nota: Não será possível selecionar **Quadrado simétrico** se a caixa de diálogo Modelo especificar condicionalidade de linha.

Criação de Medida para Ajuste de Escala Multidimensional

O ajuste de escala multidimensional utiliza dados de dissimilaridade para criar uma solução de ajuste de escala. Se seus dados forem dados multivariados (valores de variáveis medidas), deve-se criar dados de dissimilaridade para calcular uma solução de ajuste de escala multidimensional. É possível especificar os detalhes da criação de medidas de dissimilaridade a partir de seus dados.

Medida. Permite especificar a medida de dissimilaridade para sua análise. Selecione uma alternativa do grupo Medida correspondente ao seu tipo de dados e, em seguida, escolha uma das medidas na lista suspensa correspondente a esse tipo de medida. As alternativas disponíveis são:

- **Intervalo**. Distância Euclidiana, distância Euclidiana Quadrada, Chebychev, Bloco, Minkowski ou Customizado.
- **Contagens**. Medida Qui-quadrado ou medida Fi-quadrado.
- **Binário**. Distância euclidiana, distância Euclidiana Quadrática, diferença de Tamanho, diferença de Padrão, Variância ou Lance e Williams.

Criar Matriz de Distância. Permite escolher a unidade de análise. As alternativas são variáveis Between ou casos Between.

Transformar Valores. Em alguns casos, como quando variáveis são medidas em escalas muito diferentes, talvez você queira padronizar os valores antes de calcular proximidades (não aplicável a dados binários). Escolha um método de padronização na lista suspensa Padronizar. Se nenhuma padronização for necessária, escolha **Nenhum**.

Modelo de Ajuste de Escala Multidimensional

A estimação de correção de um modelo de ajuste de escala multidimensional depende dos aspectos dos dados e do próprio modelo.

Nível de Medição. Permite especificar o nível de seus dados. As alternativas são Ordinal, Intervalo ou Razão. Se as variáveis forem ordinais, selecionar **Desempatar observações empatadas** solicita que as variáveis sejam tratadas como variáveis contínuas, de modo que os empates (valores iguais para casos diferentes) sejam resolvidos corretamente.

Condicionalidade. Permite especificar quais comparações são significativas. As alternativas são Matriz, Linha ou Incondicional.

Dimensões. Permite especificar a dimensionalidade de uma ou mais soluções de ajuste de escala. Uma solução é calculada para cada número no intervalo. Especifique números inteiros entre 1 e 6; um mínimo de 1 é permitido apenas se você selecionar **Distância euclidiana** como o modelo de ajuste de escala. Para uma solução única, especifique o mesmo número para mínimo e máximo.

Modelo de Ajuste de Escala. Permite especificar as suposições pelas quais o ajuste de escala é executado. As alternativas disponíveis são distância euclidiana ou distância euclidiana de diferenças individuais (também conhecida como INDSCAL). Para o modelo de distância euclidiana de diferenças individuais, é possível selecionar **Permitir ponderações de assunto negativas**, se apropriado para seus dados.

Opções de Ajuste de Escala Multidimensional

É possível especificar opções para sua análise de ajuste de escala multidimensional.

Exibição. Permite selecionar vários tipos de saída. As opções disponíveis são gráficos de Grupo, gráficos das observações individuais, Matriz de dados e Sumarização de modelo e opções.

Critérios. Permite determinar quando a iteração deve parar. Para alterar os padrões, insira valores para **Convergência de estresse de S**, **Valor mínimo de estresse de S** e **Máximo de iterações**.

Tratar distâncias menores que n como omissas. Distâncias que forem menores que este valor são excluídas da análise.

Recursos Adicionais do Comando ALSCAL

O idioma da sintaxe de comando também permite:

- Utilizar três tipos de modelo adicionais, conhecidos como ASCAL, AINDS e GEMSCAL, na literatura sobre ajuste de escala multidimensional.
- Executar transformações polinomiais nos dados de intervalo e de proporções.
- Analisar similaridades (ao invés de distâncias) com dados ordinais.
- Analisar dados nominais.
- Salvar várias matrizes de coordenadas e de ponderações em arquivos e lê-los novamente para análise.
- Restringir desdobra multidimensional.

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Capítulo 32. Estatísticas de razão

O procedimento Estatísticas de Razão fornece uma lista abrangente de estatísticas básicas para descrever a razão entre duas variáveis de escala.

É possível classificar a saída por valores de uma variável de agrupamento em ordem crescente ou decrescente. O relatório de estatísticas de razão pode ser suprimido na saída, e os resultados podem ser salvos em um arquivo externo.

Exemplo. Existe uma boa uniformidade na razão entre o preço de avaliação e o preço de venda de casas em cada um dos cinco estados? Na saída, você pode ver que a distribuição de razões varia consideravelmente de estado para estado.

Estatísticas. Mediana, média, média ponderada, intervalos de confiança, coeficiente de dispersão (COD), coeficiente de variação centralizado em mediana, coeficiente de variação centralizado em média, diferencial relacionado a preços (PRD), desvio padrão, desvio absoluto médio (AAD), intervalo, valores mínimo e máximo, e o índice de concentração calculado para um intervalo ou porcentagem especificada pelo usuário dentro da razão mediana.

Considerações de Dados de Estatísticas de Razão

Dados. Utilize códigos ou sequências de caracteres numéricos para codificar variáveis de agrupamento (níveis de medição nominais ou ordinais).

Suposições. As variáveis que definem o numerador e o denominador da razão devem ser variáveis de escala que aceitam valores positivos.

Para Obter Estatísticas de Razão

1. Nos menus, escolha:

   Analisar > **Estatísticas Descritivas** > **Razão...**
2. Selecione uma variável numeradora.
3. Selecione uma variável denominadora.

Opcionalmente:

- Selecione uma variável de agrupamento e especifique a ordem dos grupos nos resultados.
- Escolha se deseja exibir os resultados no Visualizador.
- Escolha se deseja salvar os resultados em um arquivo externo para uso posterior, e especifique o nome do arquivo no qual os resultados são salvos.

Estatísticas de razão

Tendência Central. Medidas de tendência central são estatísticas que descrevem a distribuição das razões.

- **Mediana.** O valor desse número de razões menor que este valor e o número de razões maior que esse valor são os mesmos.
- **Média.** O resultado da soma das razões, dividindo o resultado pelo número total de razões.
- **Média Ponderada.** O resultado da divisão da média do numerador pela média do denominador. A média ponderada também é a média das razões ponderadas pelo denominador.
- **Intervalos de Confiança.** Exibe os intervalos de confiança para a média, mediana e média ponderada (se solicitado). Especifique um valor que seja maior ou igual a 0 e menor que 100 como o nível de confiança.

Dispersão. Estas estatísticas medem a quantia de variação ou de difusão nos valores observados.

- **AAD.** O desvio absoluto médio é o resultado da soma dos desvios absolutos das razões sobre a mediana e dividindo esse resultado pelo número total de razões.
- **COD.** O coeficiente de dispersão é o resultado de expressar o desvio absoluto médio como uma porcentagem da mediana.
- **PRD.** O diferencial relacionado a preços, também conhecido como o índice de regressividade, é o resultado da divisão da média pela média ponderada.
- **COV Centralizada em Mediana.** O coeficiente de variação centralizado em mediana é o resultado de expressar os quadrados médios raiz do desvio da mediana como uma porcentagem da mediana.
- **COV Centralizada em Média.** O coeficiente de variação centralizado em média é o resultado de expressar o desvio padrão como uma porcentagem da média.
- **Desvio padrão.** O desvio padrão é o resultado da soma dos desvios quadrados das razões sobre a média, dividindo o resultado pelo número total de razões menos um e obtendo a raiz quadrada positiva.
- **Intervalo.** O intervalo é o resultado de subtrair a razão mínima da razão máxima.
- **Mínimo.** O mínimo é a razão menor.
- **Máxima.** O máximo é a razão maior.

Índice de Concentração. O coeficiente de concentração mede a porcentagem de razões que caem em um intervalo. Ele pode ser calculado de duas maneiras diferentes:

- **Razões entre.** Aqui o intervalo é definido explicitamente ao especificar os valores alto e baixo do intervalo. Insira os valores para a proporção baixa e proporção alta e clique em **Incluir** para obter um intervalo.
- **Razões dentro.** Aqui o intervalo é definido implicitamente ao especificar a porcentagem da mediana. Insira um valor entre 0 e 100 e clique em **Incluir**. O término inferior do intervalo é igual a $(1 - 0{,}01 \times \text{valor}) \times \text{mediana}$, e o término superior é igual a $(1 + 0{,}01 \times \text{valor}) \times \text{mediana}$.

Capítulo 33. Curvas ROC

Este procedimento é uma maneira útil para avaliar o desempenho de esquemas de classificação em que existe uma variável com duas categorias pelas quais os assuntos são classificados.

Exemplo. Um banco deseja classificar corretamente quais clientes irão ou não ficar inadimplentes com seus empréstimos, de modo que métodos especiais são desenvolvidos para tomar essas decisões. As curvas ROC podem ser utilizadas para avaliar o grau de desempenho destes métodos.

Estatísticas. Área sob a curva ROC com um intervalo de confiança e pontos de coordenadas da curva ROC. Gráficos: curva ROC.

Métodos. A estimativa da área sob a curva ROC pode ser calculada de modo paramétrico ou não paramétrico usando um modelo exponencial binegativo.

Considerações de Dados da Curva ROC

Dados. As variáveis de teste são quantitativas. As variáveis de teste são geralmente compostas de probabilidades de análise discriminante ou de regressão logística, ou compostas de escores em uma escala arbitrária indicando a "intensidade de convicção" de um avaliador de que um assunto se enquadre em uma ou outra categoria. A variável de estado pode ser de qualquer tipo e indica a categoria real à qual um assunto pertence. O valor da variável de estado indica qual categoria deve ser considerada *positiva*.

Suposições. Supõe-se que números maiores na escala do avaliador representam uma crença maior de que o assunto pertence a uma categoria, ao passo que números menores na escala representam uma crença maior de que o assunto pertence à outra categoria. O usuário deve escolher qual direção é *positiva*. Também supõe-se que a categoria *real* à qual cada assunto pertence é conhecida.

Para Obter uma Curva ROC

1. Nos menus, escolha:
   Analisar > **Curva ROC ...**
2. Selecione uma ou mais variáveis de probabilidade de teste.
3. Selecione uma variável de estado.
4. Identifique o valor de *positivo* para a variável de estado.

Opções de Curva ROC

É possível especificar as seguintes opções para sua análise ROC:

Classificação. Permite especificar se o valor de corte deve ser incluído ou excluído ao criar uma classificação *positiva*. Esta configuração atualmente não tem nenhum efeito na saída.

Direção do Teste. Permite especificar a direção da escala com relação à categoria *positiva*.

Parâmetros para Erro Padrão da Área. Permite especificar o método para estimar o erro padrão da área abaixo da curva. Os métodos disponíveis são exponenciais não paramétricos e binegativos. Além disso, permite configurar o nível para o intervalo de confiança. O intervalo disponível é de 50,1% a 99,9%.

Valores omissos. Permite especificar como os valores omissos são manipulados.

Capítulo 34. Simulação

Modelos preditivos, como regressão linear, requerem um conjunto de entradas conhecidas para prever um valor de resultado ou de resposta. Em muitas aplicações no mundo real, no entanto, os valores de entradas são incertos. A simulação permite considerar a incerteza nas entradas para modelos preditivos e avaliar a probabilidade de vários resultados do modelo na presença de uma incerteza. Por exemplo, você tem um modelo de lucro que inclui o custo de materiais como uma entrada, mas há uma incerteza desse custo devido à volatilidade do mercado. É possível usar simulação para modelar essa incerteza e determinar o efeito que isso causa no lucro.

A simulação no IBM SPSS Statistics utiliza o método de Monte Carlo. Entradas incertas são modeladas com as distribuições de probabilidade (como a distribuição triangular), e valores simulados para essas entradas são gerados ao desenhar a partir dessas distribuições. Entradas cujos valores são conhecidos são mantidas fixas nos valores conhecidos. O modelo preditivo é avaliado utilizando um valor simulado para cada entrada incerta e valores fixos para as entradas conhecidas para calcular a resposta (ou respostas) do modelo. O processo é repetido muitas vezes (normalmente dezenas ou centenas de milhares de vezes), resultando em uma distribuição de valores de resposta que podem ser usados para responder perguntas de uma natureza probabilística. No contexto do IBM SPSS Statistics, cada repetição do processo gera um caso separado (registro) de dados que consiste no conjunto de valores simulados para as entradas incertas, nos valores das entradas fixas e na resposta (ou respostas) preditas do modelo.

Também é possível simular dados na ausência de um modelo preditivo ao especificar distribuições de probabilidade para variáveis que devem ser simuladas. Cada caso de dados gerado consiste no conjunto de valores simulados para as variáveis especificadas.

Para executar uma simulação, é necessário especificar detalhes como o modelo preditivo, as distribuições de probabilidade para as entradas incertas, as correlações entre essas entradas e os valores para quaisquer entradas fixas. Após ter especificado todos os detalhes para uma simulação, será possível executá-la e, opcionalmente, salvar as especificações em um arquivo de **plano de simulação**. É possível compartilhar o plano de simulação com outros usuários que, em seguida, poderão executar a simulação sem a necessidade de entender os detalhes de como ela foi criada.

Duas interfaces estão disponíveis para trabalhar com simulações. O Construtor de Simulação é uma interface avançada para usuários que estiverem projetando e executando simulações. Ele fornece o conjunto completo de capacidades para projetar uma simulação, salvar as especificações em um arquivo de plano de simulação, especificar a saída e executar a simulação. É possível construir uma simulação com base em um arquivo de modelo do IBM SPSS ou em um conjunto de equações customizadas que você define no Construtor de Simulação. Também é possível carregar um plano de simulação existente no Construtor de Simulação, modificar qualquer uma das configurações e executar a simulação e, opcionalmente, salvar o plano atualizado. Para os usuários que possuírem um plano de simulação e sobretudo desejarem executar a simulação, uma interface mais simples está disponível. Com ela é possível modificar as configurações que permitem executar a simulação sob diferentes condições, mas não fornece os recursos completos do Construtor de Simulação para projetar simulações.

Para designar uma simulação baseada em um arquivo de modelo

1. Nos menus, escolha:

   Analisar > **Simulação...**
2. Clique em **Selecionar arquivo de modelo SPSS** e clique em **Continuar**.
3. Abra o arquivo de modelo.

O arquivo de modelo é um arquivo XML que contém o modelo PMML criado a partir do IBM SPSS Statistics ou IBM SPSS Modeler. Consulte o tópico “Guia Modelo” na página 180 para obter mais informações

4. Na guia Simulação (no Construtor de Simulação), especifique as distribuições de probabilidade para entradas e valores simulados para entradas fixas. Se o conjunto de dados ativo contém dados históricos para as entradas simuladas, clique em **Ajustar todos** para determinar automaticamente a distribuição que melhor ajusta os dados a cada entrada, bem como determinar as correlações entre eles. Para cada entrada simulada que não está sendo ajustada para dados históricos, deve-se especificar explicitamente uma distribuição selecionando um tipo de distribuição e inserindo os parâmetros necessários.
5. Clique em **Executar** para executar a simulação. Por padrão, o plano de simulação, especificando os detalhes da simulação, é salvo no local especificado nas configurações de Salvar.

As seguintes opções estão disponíveis:

- Modificar o local para o plano de simulação salvo.
- Especificar correlações conhecidas entre entradas simuladas.
- Calcular automaticamente uma tabela de contingência de associações entre entradas categóricas e utilizar essas associações quando os dados forem gerados para essas entradas.
- Especificar análise de sensibilidade para investigar o efeito de variar o valor de uma entrada fixa ou de variar um parâmetro de distribuição para uma entrada simulada.
- Especificar opções avançadas como configurar o número máximo de casos para gerar ou solicitar amostragem de causa.
- Customizar saída.
- Salvar os dados simulados como um arquivo de dados.

Para designar uma simulação baseada em equações customizadas

1. Nos menus, escolha:

   Analisar > **Simulação...**
2. Clique em **Tipo em equações** e clique em **Continuar**.
3. Clique em **Nova equação** na guia Modelo (no Construtor de Simulação) para definir cada equação em seu modelo preditivo.
4. Clique na guia Simulação e especifique as distribuições de probabilidade para entradas e valores simulados para entradas fixas. Se o conjunto de dados ativo contém dados históricos para as entradas simuladas, clique em **Ajustar todos** para determinar automaticamente a distribuição que melhor ajusta os dados a cada entrada, bem como determinar as correlações entre eles. Para cada entrada simulada que não está sendo ajustada para dados históricos, deve-se especificar explicitamente uma distribuição selecionando um tipo de distribuição e inserindo os parâmetros necessários.
5. Clique em **Executar** para executar a simulação. Por padrão, o plano de simulação, especificando os detalhes da simulação, é salvo no local especificado nas configurações de Salvar.

As seguintes opções estão disponíveis:

- Modificar o local para o plano de simulação salvo.
- Especificar correlações conhecidas entre entradas simuladas.
- Calcular automaticamente uma tabela de contingência de associações entre entradas categóricas e utilizar essas associações quando os dados forem gerados para essas entradas.
- Especificar análise de sensibilidade para investigar o efeito de variar o valor de uma entrada fixa ou de variar um parâmetro de distribuição para uma entrada simulada.
- Especificar opções avançadas como configurar o número máximo de casos para gerar ou solicitar amostragem de causa.
- Customizar saída.

- Salvar os dados simulados como um arquivo de dados.

Para designar uma simulação sem um modelo preditivo

1. Nos menus, escolha:

   Analisar > **Simulação...**
2. Clique em **Criar dados simulados** e clique em **Continuar**.
3. Na guia Modelo (no Construtor de Simulação), selecione os campos que você deseja simular. É possível selecionar campos a partir do conjunto de dados ativo ou definir novos campos clicando em **Novo**.
4. Clique na guia de Simulação e especifique as distribuições de probabilidade para os campos que devem ser simulados. Se o conjunto de ativo contém dados históricos para qualquer um desses campos, clique em **Ajustar todos** para determinar automaticamente a distribuição que melhor ajusta os dados e para determinar as correlações entre os campos. Para campos que não se ajustam aos para dados históricos, deve-se especificar explicitamente uma distribuição, selecionando um tipo de distribuição e inserindo os parâmetros necessários.
5. Clique em **Executar** para executar a simulação. Por padrão, os dados simulados são salvos no novo conjunto de dados especificados nas configurações de Salvar. Além disso, o plano de simulação, que especifica os detalhes da simulação, é salvo no local especificado nas configurações de Salvar.

As seguintes opções estão disponíveis:

- Modifique o local para os dados simulados ou o plano de simulação salvo.
- Especifique as correlações conhecidas entre campos simulados conhecidos.
- Calcule automaticamente uma tabela de contingência de associações entre campos categóricos e use essas associações quando os dados são gerados para esses campos.
- Especifique a análise de sensibilidade para investigar o efeito de variar um parâmetro de distribuição para um campo simulado.
- Especifique opções avançadas, como definir o número de casos para gerar.

Para executar uma simulação a partir de um plano de simulação

Duas opções estão disponíveis para executar uma simulação a partir de um plano de simulação. É possível usar o diálogo Executar Simulação, que é designado principalmente para executar a partir de um plano de simulação, ou é possível usar o Construtor de Simulação.

Para usar o diálogo Executar Simulação:

1. Nos menus, escolha:

   Analisar > **Simulação...**
2. Clique em **Abrir um plano de simulação existente**.
3. Certifique-se de que caixa de seleção **Abrir no Construtor de Simulação** não esteja marcada e clique em **Continuar**.
4. Abra o plano de simulação.
5. Clique em **Executar** no diálogo Executar Simulação.

Para executar a simulação a partir do Construtor de Simulação:

1. Nos menus, escolha:

   Analisar > **Simulação...**
2. Clique em **Abrir um plano de simulação existente**.
3. Marque a caixa de seleção **Abrir no Construtor de Simulação** e clique em **Continuar**.
4. Abra o plano de simulação.

5. Modifique quaisquer configurações que você deseja modificar na guia Simulação.
6. Clique em **Executar** para executar a simulação.

Opcionalmente, é possível fazer o seguinte:

- Configurar ou modificar análise de sensibilidade para investigar o efeito de variar o valor de uma entrada fixa ou variar um parâmetro de distribuição para uma entrada simulada.
- Reajuste as distribuições e correlações para entradas simuladas para novos dados.
- Altere a distribuição para uma entrada simulada.
- Customizar saída.
- Salve os dados simulados em um arquivo de dados.

Construtor de Simulação

O Construtor de Simulação fornece o conjunto completo de capacidades para projetar e executar simulações. Ele permite executar as seguintes tarefas gerais:

- Projetar e executar uma simulação para um modelo do IBM SPSS definido em um arquivo de modelo PMML.
- Projetar e executar uma simulação para um modelo preditivo definido por um conjunto de equações customizadas que você especificar.
- Projetar e executar uma simulação que gera dados na ausência de um modelo preditivo.
- Executar uma simulação com base em um plano de simulação existente, opcionalmente modificando quaisquer configurações do plano.

Guia Modelo

Para simulações com base em um modelo preditivo, a guia Modelo especifica a origem do modelo. Para simulações que não incluem um modelo preditivo, a guia Modelo especifica os campos que devem ser simulados.

Selecione um arquivo de modelo SPSS. Esta opção especifica que o modelo preditivo é definido em um arquivo de modelo do IBM SPSS. Um arquivo de modelo do IBM SPSS é um arquivo XML ou um archive compactado (arquivo .zip) que contém o modelo PMML criado a partir do IBM SPSS Statistics ou do IBM SPSS Modeler. Os modelos preditivos são criados por procedimentos, como Regressão Linear e Árvores de Decisão, no IBM SPSS Statistics e podem ser exportados em um arquivo de modelo. É possível utilizar um arquivo de modelo diferente clicando em **Procurar** e navegando para o arquivo desejado.

Modelos PMML suportados por Simulação

- Regressão linear
- Modelo Linear Automático
- Modelo linear generalizado
- Modelo Linear Generalizado Misto
- Modelo linear geral
- Regressão logística binária
- Regressão logística multinomial
- Regressão Multnomial Ordinal
- Regressão de Cox
- Árvore
- Árvore Impulsionada (C5)
- Discriminante
- Clusterização em duas etapas

- Cluster por K-médias
- Rede Neural
- Conjunto de regras (lista de decisão)

Nota:
- Os modelos PMML que possuem diversos campos de destino (variáveis) ou divisões não são suportados para uso em Simulação.
- Os valores das entradas de sequência de caracteres para modelos de regressão logística binários estão limitados a 8 bytes no modelo. Se estiver ajustando essas entradas de sequência de caracteres para o conjunto de dados ativo, certifique-se de que os valores nos dados não excedam 8 bytes de comprimento. Os valores de dados que excederem 8 bytes são excluídos da distribuição categórica associada para a entrada e são exibidos como não correspondentes na tabela de saída Categorias Não Correspondentes.

Tipo nas equações para o modelo. Esta opção especifica que o modelo preditivo consiste em uma ou mais equações customizadas a serem criadas por você. Crie equações clicando em **Nova Equação**. Isto abre o Editor de Equação. É possível modificar equações existentes, copiá-las para utilizar como modelos para novas equações, reordená-las e excluí-las.
- O Construtor de Simulação não suporta sistemas de equações simultâneas ou equações que não são lineares na variável de destino.
- As equações customizadas são avaliadas na ordem em que elas são especificadas. Se a equação para uma determinada resposta depender de outra resposta, então a outra resposta deverá ser definida por uma equação anterior.

  Por exemplo, dado o conjunto de três equações a seguir, a equação para *profit* depende dos valores de *revenue* e *expenses*, de modo que as equações para *revenue* e *expenses* devem preceder a equação para *profit*.

```
revenue = price*volume
expenses = fixed + volume*(unit_cost_materials + unit_cost_labor)
profit = revenue - expenses
```

Criar dados simulados sem um modelo. Selecione esta opção para simular dados sem um modelo preditivo. Especifique os campos que devem ser simulados selecionando campos do conjunto de dados ativo ou clicando em **Novo** para definir novos campos.

Editor de Equação

O Editor de Equação permite criar ou modificar uma equação customizada para seu modelo preditivo.
- A expressão para a equação pode conter campos do conjunto de dados ativo ou novos campos de entrada que você define no Editor de Equação.
- É possível especificar propriedades da resposta como seu nível de medição, rótulos de valor e se a saída é gerada para a resposta.
- É possível utilizar respostas de equações definidas anteriormente como entradas para a equação atual, permitindo criar equações acopladas.
- É possível anexar um comentário descritivo na equação. Os comentários são exibidos junto da equação na guia Modelo.

1. Insira o nome da resposta. Opcionalmente, clique em **Editar** na caixa de texto Resposta para abrir o diálogo Entradas Definidas, permitindo alterar as propriedades padrão da resposta.
2. Para construir uma expressão, cole componentes no campo Expressão Numérica ou digite diretamente no campo Expressão Numérica.

- É possível construir sua expressão utilizando campos do conjunto de dados ativo ou definir novas entradas clicando no botão **Novo**. Isso abre o diálogo Definir Entradas.
- É possível colar funções selecionando um grupo da lista de grupos Função e dando um clique duplo na função na lista Funções (ou selecione a função e clique na seta adjacente à lista de grupos Função).

Insira quaisquer parâmetros indicados por pontos de interrogação. O grupo de funções rotulado como **Tudo** fornece uma lista de todas as funções disponíveis. Uma breve descrição da função selecionada atualmente é exibida em uma área reservada na caixa de diálogo.

- Constantes da sequência devem ser colocadas entre aspas.
- Se os valores contiverem decimais, (.) deve ser usado como o indicador decimal.

Nota: A simulação não suporta equações customizadas com respostas de sequência de caracteres.

Entradas Definidas: O diálogo Entradas Definidas permite definir novas entradas e configurar propriedades para respostas.

- Se uma entrada a ser utilizada em uma equação não existir no conjunto de dados ativos, ela deverá ser definida antes de poder ser utilizada na equação.
- Se estiver simulando dados sem um modelo preditivo, deve-se definir todas as entradas simuladas que não existirem no conjunto de dados ativo.

Nome. Especifique o nome para uma resposta ou entrada.

Resposta. Também é possível especificar um nível de medição para uma resposta. O nível de medição padrão é contínuo. Também é possível especificar se a saída será criada para esta resposta. Por exemplo, para um conjunto de equações acopladas, talvez você esteja interessado somente na saída da resposta para a equação final, de modo que você suprime a saída das outras respostas.

Entrada a ser simulada. Isso especifica que os valores da entrada serão simulados de acordo com uma distribuição de probabilidade especificada (a distribuição de probabilidade é especificada na guia Simulação). O nível de medição determina o conjunto padrão de distribuições que são consideradas ao localizar a distribuição que melhor ajusta os dados para a entrada (ao clicar em **Ajustar** ou **Ajustar Tudo** na guia Simulação). Por exemplo, se o nível de medição for contínuo, então a distribuição normal (apropriada para dados contínuos) será considerada, mas a distribuição binomial não.

Nota: Selecione um nível de medição de String para entradas de sequência de caracteres. As entradas de sequência que devem ser simuladas são restritas à Distribuição Categórica.

Entrada de valor fixo. Isso especifica que o valor da entrada é conhecido e será fixo no valor conhecido. Entradas fixas podem ser numéricas ou sequência de caracteres. Especifique um valor para a entrada fixa. Os valores de sequência não devem ser colocados entre aspas.

Rótulos de valor. É possível especificar rótulos de valor para respostas, entradas simuladas e entradas fixas. Rótulos de valor são utilizados nos gráficos e tabelas de saída.

Guia Simulação

A guia Simulação especifica todas as propriedades da simulação que são diferentes do modelo preditivo. É possível executar as seguintes tarefas gerais na guia Simulação:

- Especificar distribuições de probabilidade para as entradas simuladas e valores para entradas fixas.
- Especificar as correlações entre as entradas simuladas. Para as entradas categóricas, é possível especificar que as associações que existem entre essas entradas no conjunto de dados ativo são usadas quando os dados são gerados para essas entradas.
- Especificar opções avançadas, como a amostragem de rodapé e os critérios para ajustar as distribuições para dados históricos.
- Customizar saída.
- Especificar onde salvar o plano de simulação e, opcionalmente, salvar os dados simulados.

Campos Simulados

Para executar uma simulação, cada campo de entrada deve ser especificado como fixo ou simulado. Entradas simuladas são aquelas cujos valores são incertos e serão geradas ao desenhar a partir de uma distribuição de probabilidade especificada. Quando os dados históricos estão disponíveis para as entradas a serem simuladas, as distribuições que melhor se ajustarem aos dados podem ser determinadas automaticamente, junto de quaisquer correlações entre essas entradas. Também é possível especificar manualmente distribuições ou correlações se os dados históricos não estiverem disponíveis ou se você precisar de distribuições ou correlações específicas.

Entradas fixas são aquelas cujos valores são conhecidos e que permanecem constantes para cada caso gerado na simulação. Por exemplo, você possui um modelo de regressão linear para vendas como uma função de um número de entradas que incluem preço, e você deseja manter o preço fixo ao preço de mercado atual. Você então especifica o preço como uma entrada fixa.

Para simulações com base em um modelo preditivo, cada um preditor no modelo é um campo de entrada para a simulação. Para simulações que não incluem um modelo preditivo, os campos que são especificados na guia Modelo são as entradas para a simulação.

Ajustando automaticamente as distribuições e calculando as correlações para as entradas simuladas. Se o conjunto de dados ativo contiver dados históricos para as entradas que deseja simular, então será possível localizar automaticamente as distribuições que melhor ajustam os dados para essas entradas, além de determinar quaisquer correlações entre elas. As etapas são as seguintes:

1. Verifique se cada uma das entradas que deseja simular é correspondida com o campo correto no conjunto de dados ativo. As entradas são listadas na coluna de entrada e a coluna Ajustar a exibe o campo correspondente no conjunto de dados ativo. É possível corresponder uma entrada a um campo diferente no conjunto de dados ativo ao selecionar um item diferente na lista suspensa Ajustar a.

   Um valor de *-None-* na coluna Ajustar a indica que a entrada não pôde ser correspondida automaticamente a um campo no conjunto de dados ativo. Por padrão, as entradas são correspondidas aos campos de conjunto de dados por nome, nível de medição e tipo (numérico ou sequência de caracteres). Se o conjunto de dados ativo não contiver dados históricos para a entrada, então especifique manualmente a distribuição para a entrada ou especifique a entrada como uma entrada fixa, conforme descrito abaixo.
2. Clique em **Ajustar Tudo**.

A distribuição de ajuste mais próxima e seus parâmetros associados são exibidos na coluna Distribuição junto de um gráfico da distribuição sobreposta em um histograma (ou gráfico de barras) dos dados históricos. As correlações entre entradas simuladas são exibidas nas configurações de Correlações. É possível examinar os resultados do ajuste e customizar o ajuste de distribuição automático para uma entrada específica ao selecionar a linha para a entrada e clicar em **Detalhes do Ajuste**. Consulte o tópico “Detalhes do Ajuste” na página 185 para obter mais informações

É possível executar ajuste de distribuição automático para uma entrada específica ao selecionar a linha para a entrada e clicar em **Ajustar**. Correlações para todas as entradas simuladas que correspondem aos campos no conjunto de dados ativo também são calculadas automaticamente.

Nota:

- Casos com valores omissos para qualquer entrada simulada são excluídos do ajuste de distribuição, do cálculo de correlações e do cálculo da tabela de contingência opcional (para entradas com uma distribuição Categórica). Opcionalmente, é possível especificar se valores omissos de usuário de entradas com uma Distribuição Categórica são tratados como válidos. Por padrão, eles são tratados como omissos. Para obter informações adicionais, consulte o tópico “Opções Avançadas” na página 187.
- Para entradas contínuas e ordinais, se um ajuste aceitável não puder ser localizado para nenhuma das distribuições testadas, então a distribuição Empírica será sugerida como o ajuste mais próximo. Para entradas contínuas, a distribuição Empírica é a função de distribuição cumulativa dos dados históricos. Para entradas ordinais, a distribuição Empírica é a distribuição categórica dos dados históricos.

Especificando distribuições manualmente. É possível especificar manualmente a distribuição de probabilidade para qualquer entrada simulada ao selecionar a distribuição na lista suspensa **Tipo** e inserir os parâmetros de distribuição na grade Parâmetros. Após ter inserido os parâmetros para uma distribuição, um gráfico de amostra da distribuição, com base nos parâmetros especificados, será exibido ao lado da grade Parâmetros. A seguir há algumas notas sobre distribuições específicas:

- **Categóricos.** A distribuição categórica descreve um campo de entrada que possui um número fixo de valores, referidos como categorias. Cada categoria possui uma probabilidade associada, de forma que a soma das probabilidades de todas as categorias seja igual a um. Para inserir uma categoria, clique na coluna à esquerda na grade Parâmetros e especifique o valor de categoria. Insira a probabilidade associada à categoria na coluna à direita.

  Nota: Entradas categóricas a partir de um modelo PMML possuem categorias que são determinadas a partir do modelo e não podem ser modificadas.
- **Binomial Negativa - Falhas.** Descreve a distribuição do número de falhas em uma sequência de avaliações antes que um número especificado de sucessos seja observado. O parâmetro *thresh* é o número especificado de sucessos e o parâmetro *prob* é a probabilidade de sucesso em qualquer avaliação determinada.
- **Binomial Negativa - Avaliações.** Descreve a distribuição do número de avaliações necessárias antes que um número especificado de sucessos seja observado. O parâmetro *thresh* é o número especificado de sucessos e o parâmetro *prob* é a probabilidade de sucesso em qualquer avaliação determinada.
- **Intervalo.** Essa distribuição consiste em um conjunto de intervalos com uma probabilidade designada a cada intervalo, de modo que a soma das probabilidades de todos os intervalos seja igual a 1. Os valores dentro de um intervalo especificado são obtidos a partir de uma distribuição uniforme definida nesse intervalo. Os intervalos são especificados ao inserir um valor mínimo, um valor máximo e uma probabilidade associada.

  Por exemplo, você acredita que o custo de uma matéria-prima tem 40% de chance de cair no intervalo de $10 - $15 por unidade e 60% de chance de cair no intervalo de $15 -$20 por unidade. Você modela o custo com uma distribuição Intervalo consistindo em dois intervalos [10 – 15] e [15 – 20], configurando a probabilidade associada ao primeiro intervalo para 0,4 e a probabilidade associada ao segundo intervalo para 0,6. Os intervalos não precisam ser contínuos e podem até mesmo ser sobrepostos. Por exemplo, você pode ter especificado os intervalos de $10 - $15 e $20 - $25 ou $10 - $15 e $13 - $16.
- **Weibull.** O parâmetro *c* é um parâmetro de localização opcional, que especifica onde a origem da distribuição está localizada.

Os parâmetros para as seguintes distribuições possuem o mesmo significado que as funções de variável aleatórias associadas disponíveis na caixa de diálogo Calcular Variável: Bernoulli, beta, binomial, exponencial, gama, lognormal, binomial negativa (avaliações e falhas), normal, Poisson e uniforme.

Especificando entradas fixas. Especifique uma entrada fixa ao selecionar Fixo a partir da lista suspensa **Tipo** na coluna Distribuição e inserir o valor fixo. O valor pode ser numérico ou sequência de caracteres, dependendo se a entrada é numérica ou uma sequência de caracteres. Os valores de sequência não devem ser colocados entre aspas.

Especificando limites em valores simulados. A maioria das distribuições suporta a especificação de limites superior e inferior em valores simulados. É possível especificar um limite inferior ao inserir um valor na caixa de texto **Mín.**, e especificar um limite superior ao inserir um valor na caixa de texto **Máx.**.

Bloqueando entradas. Bloquear uma entrada, ao marcar a caixa de seleção na coluna com o ícone de bloqueio, exclui a entrada do ajuste distribuição automático. Isso é mais útil ao especificar manualmente uma distribuição ou um valor fixo e desejar assegurar que essa entrada não seja afetada pelo ajuste de distribuição automático. O bloqueio também será útil se pretender compartilhar seu plano de simulação com usuários que irão executá-lo no diálogo Executar Simulação e desejar impedir quaisquer mudanças em determinadas entradas. Neste sentido, as especificações para entradas bloqueadas não podem ser modificadas no diálogo Executar Simulação.

Análise de Sensibilidade. A análise de sensibilidade permite investigar o efeito das mudanças sistemáticas em uma entrada fixa ou em um parâmetro de distribuição para uma entrada simulada ao gerar um conjunto independente de casos simulados – efetivamente, uma simulação separada - para cada valor especificado. Para especificar análise de sensibilidade, selecione uma entrada fixa ou simulada e clique em **Análise de Sensibilidade**. A análise de sensibilidade é limitada a uma entrada fixa única ou a um parâmetro de distribuição único para uma entrada simulada. Consulte o tópico "Análise de Sensibilidade" na página 186 para obter informações adicionais.

Ícone de status de Ajuste

Os ícones na coluna Ajustar a indicam o status de ajuste para cada campo de entrada.

Tabela 3. Ícones de status.

Ícone	Descrição
	Nenhuma distribuição foi especificada para a entrada, e a entrada não foi especificada como fixa. Para executar a simulação, deve-se especificar uma distribuição para esta entrada ou defini-la para ser fixa e especificar o valor fixo.
	A entrada foi ajustada anteriormente para um campo que não existe no conjunto de dados ativo. Nenhuma ação é necessária, a menos que você deseje reajustar a distribuição para a entrada no conjunto de dados ativo.
	A distribuição de ajuste mais próxima foi substituída por uma distribuição alternativa no diálogo Detalhes do Ajuste.
	A entrada é configurada como a distribuição de ajuste mais próxima.
	A distribuição foi especificada manualmente ou as iterações da análise de sensibilidade foram especificadas para essa entrada.

Detalhes do Ajuste: O diálogo Detalhes do Ajuste exibe os resultados do ajuste de distribuição automático para uma entrada específica. As distribuições são ordenadas pela Qualidade do ajuste, com a distribuição de ajuste mais próxima listada primeiro. É possível substituir a distribuição de ajuste mais próxima ao selecionar o botão de opções para a distribuição desejada na coluna Usar. Selecionar um botão de opções na coluna Usar também exibe um gráfico da distribuição sobreposto em um histograma (ou um gráfico de barras) dos dados históricos para essa entrada.

Estatística de ajuste. Por padrão e para campos contínuos, o teste de Anderson-Darling é utilizado para determinar a Qualidade do ajuste. Como alternativa e somente para campos contínuos, é possível especificar o teste de Kolmogorov-Smirnoff para Qualidade do ajuste ao selecionar essa opção nas configurações de Opções Avançadas. Para entradas contínuas, os resultados dos dois testes são mostrados na coluna Estatísticas de Ajuste (A para Anderson-Darling e K para Kolmogorov-Smirnoff), com o teste escolhido utilizado para ordenar as distribuições. Para entradas ordinais e nominais, o teste qui-quadrado é utilizado. Os valores p associados aos testes também são mostrados.

Parâmetros. Os parâmetros de distribuição associados a cada distribuição ajustada são exibidos na coluna Parâmetros. Os parâmetros para as seguintes distribuições possuem o mesmo significado que as funções de variável aleatórias associadas disponíveis na caixa de diálogo Calcular Variável: Bernoulli, beta, binomial, exponencial, gama, lognormal, binomial negativa (avaliações e falhas), normal, Poisson e uniforme. Consulte o tópico para obter mais informações Para a distribuição categórica, os nomes dos parâmetros são as categorias e os valores dos parâmetros são as probabilidades associadas.

Reajustando com um conjunto de distribuição customizado. Por padrão, o nível de medição da entrada é utilizado para determinar o conjunto de distribuições consideradas para ajuste de distribuição

automático. Por exemplo, distribuições contínuas como lognormal e gama são consideradas ao ajustar uma entrada contínua, mas distribuições discretas, como Poisson e binomial, não são. É possível escolher um subconjunto das distribuições padrão ao selecionar as distribuições na coluna Reajustar. Também é possível substituir o conjunto padrão das distribuições ao selecionando um nível de medição diferente da lista suspensa **Tratar como (Medida)** e selecionar as distribuições na coluna Reajustar. Clique em **Executar Reajuste** para reajustar com o conjunto de distribuição customizado.

Nota:

- Casos com valores omissos para qualquer entrada simulada são excluídos do ajuste de distribuição, do cálculo de correlações e do cálculo da tabela de contingência opcional (para entradas com uma distribuição Categórica). Opcionalmente, é possível especificar se valores omissos de usuário de entradas com uma Distribuição Categórica são tratados como válidos. Por padrão, eles são tratados como omissos. Para obter informações adicionais, consulte o tópico "Opções Avançadas" na página 187.
- Para entradas contínuas e ordinais, se um ajuste aceitável não puder ser localizado para nenhuma das distribuições testadas, então a distribuição Empírica será sugerida como o ajuste mais próximo. Para entradas contínuas, a distribuição Empírica é a função de distribuição cumulativa dos dados históricos. Para entradas ordinais, a distribuição Empírica é a distribuição categórica dos dados históricos.

Análise de Sensibilidade: A análise de sensibilidade permite investigar o efeito de variar uma entrada fixa ou um parâmetro de distribuição para uma entrada simulada através de um conjunto de valores especificado. Um conjunto de casos simulados independentes - efetivamente, uma simulação separada - é gerado para cada valor especificado, permitindo investigar o efeito de variar a entrada. Cada conjunto de casos simulados é referido como uma **iteração**.

Iterar. Esta opção permite especificar o conjunto de valores sobre qual entrada será variada.

- Se estiver variando o valor de um parâmetro de distribuição, selecione o parâmetro na lista suspensa. Insira o conjunto de valores no valor de Parâmetro pela grade de iteração. Clicar em **Continuar** incluirá os valores especificados na grade Parâmetros da entrada associada, com um índice especificando o número de iteração do valor.
- Para distribuições Categóricas e de Intervalo, as probabilidades das categorias ou intervalos podem ser respectivamente variadas, mas os valores das categorias e dos terminais dos intervalos não podem ser variados. Selecione uma categoria ou intervalo na lista suspensa e especifique o conjunto de probabilidades no valor de Parâmetro por grade de iteração. As probabilidades para as outras categorias ou intervalos serão ajustadas de modo automático e apropriado.

Nenhuma iteração. Utilize essa opção para cancelar as iterações para uma entrada. Clicar em **Continuar** removerá as iterações.

Correlações

Os campos de entrada a serem simulados geralmente são conhecidos como estando correlacionados - por exemplo, altura e peso. As correlações entre entradas que serão simuladas devem ser levadas em conta a fim de assegurar que os valores simulados preservam essas correlações.

Recalcular correlações ao ajustar. Essa opção especifica que as correlações entre entradas simuladas são calculadas automaticamente ao ajustar as distribuições para o conjunto de dados ativo por meio das ações **Ajustar Tudo** ou **Ajustar** nas configurações de Campos Simulados.

Não recalcular correlações ao ajustar. Selecione essa opção se desejar especificar as correlações manualmente e evitar que elas sejam sobrescritas ao ajustar automaticamente as distribuições para o conjunto de dados ativo. Os valores que forem inseridos na grade de Correlações devem estar entre -1 e 1. Um valor 0 especifica que não há nenhuma correlação entre o par de entradas associado.

Reconfigurar. Isso reconfigura todas as correlações para 0.

Use a tabela de contingência multiponto ajustada para entradas com distribuição categórica. Para entradas com distribuição categórica, é possível calcular automaticamente uma tabela de contingência multiponto a partir do conjunto de dados ativo que descreve as associações entre essas entradas. Em seguida, a tabela de contingência é utilizada quando os dados são gerados para essas entradas. Se optar por salvar o plano de simulação, a tabela de contingência é salva no arquivo de plano e utilizada ao executar o plano.

- **Calcular a tabela de contingência a partir do conjunto de dados ativo.** Se estiver trabalhando com um plano de simulação existente que contenha uma tabela de contingência, será possível recalcular a tabela de contingência a partir do conjunto de dados ativo. Esta ação substitui a tabela de contingência no arquivo de plano carregado.
- **Utilizar tabela de contingência a partir do plano de simulação carregado.** Por padrão, ao carregar um plano de simulação que contenha uma tabela de contingência, a tabela do plano é utilizada. É possível recalcular a tabela de contingência a partir do conjunto de ativo ao selecionar **Calcular tabela de contingência a partir do conjunto de dados ativo**.

Opções Avançadas

Número Máximo de Casos. Isso especifica o número máximo de casos de dados simulados e os valores de resposta associados a serem gerados. Quando a análise de sensibilidade é especificada, este é o número máximo de casos para cada iteração.

Resposta para critérios de parada. Se o seu modelo preditivo contiver mais de uma resposta, será possível selecionar a resposta para a qual os critérios de parada são aplicados.

Critérios de parada. Essas opções especificam critérios para parar a simulação, potencialmente antes de o número máximo de casos permitidos ter sido gerado.

- **Continuar até que o máximo seja atingido.** Isso especifica que casos simulados serão gerados até que o número máximo de casos seja atingido.
- **Parar quando as caudas tiverem sido amostradas.** Utilize esta opção quando desejar assegurar que uma das caudas de uma distribuição de resposta especificada foi corretamente amostrada. Os casos simulados serão gerados até que a amostragem de cauda especificada esteja concluída ou o número máximo de casos seja atingido. Se o seu modelo preditivo contiver diversas respostas, então selecione a resposta à qual estes critérios serão aplicados, na lista suspensa **Resposta para critérios de parada**.

  Tipo. É possível definir o limite da região da cauda ao especificar um valor da resposta, como 10.000.000 ou um percentil, como o 99º percentil. Se escolher Valor na lista suspensa **Tipo**, insira o valor do limite na caixa de texto Valor e utilize a lista suspensa **Lado** para especificar se este é o limite da região da cauda Esquerda ou da região da cauda Direita. Se escolher Percentil na lista suspensa **Tipo**, insira um valor na caixa de texto Percentil.

  Frequência. Especifique o número de valores da resposta que devem estar na região da cauda para assegurar que a cauda foi corretamente amostrada. A geração de casos será interrompida quando esse número for atingido.
- **Parar quando o intervalo de confiança da média estiver dentro do limite especificado.** Utilize essa opção quando desejar assegurar que a média de uma determinada resposta seja conhecida com um grau de precisão especificado. Os casos simulados serão gerados até que o grau de precisão especificado tenha sido atingido ou o número máximo de casos seja atingido. Para utilizar essa opção, especifique um nível de confiança e um limite. Os casos simulados serão gerados até que o intervalo de confiança associado ao nível especificado esteja dentro do limite. Por exemplo, é possível utilizar essa opção para especificar que casos são gerados até que o intervalo de confiança da média no nível de confiança de 95% esteja dentro de 5% do valor médio. Se o seu modelo preditivo contiver diversas respostas, então selecione a resposta à qual estes critérios serão aplicados, na lista suspensa **Resposta para critérios de parada**.

  Tipo de Limite. É possível especificar o limite como o valor numérico ou como um percentual da média. Se escolher Valor na lista suspensa **Tipo de Limite**, insira um limite na caixa de texto Limite como Valor. Se escolher Porcentagem na lista suspensa **Tipo do Limite**, insira um valor na caixa de texto Limite como Porcentagem.

Número de casos para amostra. Isso especifica o número de casos a serem utilizados ao ajustar automaticamente as distribuições de entradas simuladas no conjunto de dados ativo. Se o seu conjunto de dados for muito grande, talvez você queira considerar limitar o número de casos utilizados para ajuste de distribuição. Se selecionar **Limitar para N casos**, os primeiros N casos serão utilizados.

Critérios de Qualidade do ajuste (Contínuos). Para entradas contínuas, é possível utilizar o teste Anderson-Darling ou o teste de qualidade do ajuste de Kolmogorov-Smirnoff para classificar as distribuições ao ajustar as distribuições de entradas simuladas no conjunto de ativo. O teste de Anderson-Darling é selecionado por padrão e é recomendado especialmente quando desejar assegurar o melhor ajuste possível nas regiões de cauda.

Distribuição Empírica. Para entradas contínuas, a distribuição Empírica é a função de distribuição cumulativa dos dados históricos. É possível especificar o número de categorias utilizadas para calcular a distribuição Empírica para entradas contínuas. O padrão é 100 e o máximo é 1000.

Replicar resultados. Configurar uma semente aleatória permite replicar sua simulação. Especifique um número inteiro ou clique em **Gerar**, que criará um número inteiro pseudoaleatório entre 1 e 2147483647, inclusive. O padrão é 629111597.

Nota: Para uma determinada semente aleatória, resultados são replicados a menos que o número de encadeamentos seja alterado. Em um computador específico, o número de encadeamentos é constante a menos que o mude executando a sintaxe de comando `SET THREADS`. O número de encadeamentos pode mudar se você executar a simulação em um computador diferente devido a um algoritmo interno ser usado para determinar o número de encadeamentos para cada computador.

Valores omissos do usuário para entradas com uma Distribuição Categórica. Esses controles especificam se valores omissos de usuário de entradas com uma Distribuição Categórica são tratados como válidos. Valores omissos de sistema e valores omissos de usuário para todos os outros tipos de entradas são sempre tratados como inválidos. Todas as entradas devem ter valores válidos para um caso a ser incluído no ajuste de distribuição, no cálculo de correlações e no cálculo da tabela de contingência opcional.

Funções de Densidade

Essas configurações permitem customizar a saída para funções de densidade de probabilidade e funções de distribuição cumulativas para variáveis resposta contínuas, bem como gráficos de barras de valores preditos para variáveis resposta categóricas.

Função de Densidade de Probabilidade (PDF). A função de densidade de probabilidade exibe a distribuição de valores de resposta. Para variáveis resposta contínuas, ela permite determinar a probabilidade de que a resposta esteja dentro de uma determinada região. Para variáveis resposta categóricas (respostas com um nível de medição de nominal ou ordinal), um gráfico de barras é gerado, exibindo a porcentagem de casos que caem em cada categoria da resposta. Opções adicionais para variáveis resposta categóricas de modelos PMML estão disponíveis com os valores de Categoria para relatar a configuração descrita a seguir.

Para modelos de cluster Two-Step e modelos de cluster K-Médias, um gráfico de barras de associação de cluster é produzido.

Função de Distribuição Cumulativa (CDF). A função de distribuição cumulativa exibe a probabilidade de que o valor da resposta seja menor ou igual a um valor especificado. Ela está disponível apenas para variáveis resposta contínuas.

Posições da régua de controle. É possível especificar as posições iniciais das linhas de referência móveis em gráficos PDF e CDF. Valores que são especificados para as linhas inferior e superior referem-se a posições ao longo do eixo horizontal, e não a percentis. É possível remover a linha inferior ao selecionar **-Infinito** ou remover a linha superior ao selecionar **Infinito**. Por padrão, as linhas são posicionadas nos 5º e 95º percentis. Quando diversas funções de distribuição são exibidas em um único gráfico (devido a

diversas respostas ou resultados das iterações de análise de sensibilidade), o padrão refere-se à distribuição da primeira iteração ou da primeira resposta.

Linhas de Referência (Contínuas). É possível solicitar várias linhas de referência verticais a serem incluídas nas funções de densidade de probabilidade e nas funções de distribuição cumulativas para variáveis resposta contínuas.

- **Sigmas.** É possível incluir linhas de referência em mais e menos um número especificado de desvios padrão da média de uma resposta.
- **Percentis.** É possível incluir linhas de referência em um ou dois valores de percentil da distribuição de uma resposta ao inserir valores nas caixas de texto Inferior e Superior. Por exemplo, um valor de 95 na caixa de texto Superior representa o 95⌂ percentil, que é o valor abaixo do qual 95% das observações caem. Da mesma forma, um valor de 5 na caixa de texto Inferior representa o 5⌂ percentil, que é o valor abaixo do qual 5% das observações caem.
- **Linhas de referência customizadas.** É possível incluir linhas de referência com valores especificados da resposta.

Nota: Quando diversas funções de distribuição são exibidas em um único gráfico (devido a diversas respostas ou resultados das iterações de análise de sensibilidade), as linhas de referência são aplicadas somente à distribuição da primeira iteração ou da primeira resposta. É possível incluir linhas de referência às outras distribuições do diálogo Opções de Gráfico, que é acessado a partir do gráfico CDF ou PDF.

Sobrepor resultados a partir de variáveis resposta contínuas separadas. No caso de diversas variáveis resposta contínuas, isso especifica se as funções de distribuição de todas as respostas são exibidas em um único gráfico, com um gráfico para funções de densidade de probabilidade e outro para funções de distribuição cumulativas. Quando essa opção não é selecionada, os resultados de cada resposta são exibidos em um gráfico separado.

Valores de categoria para relatório. Para modelos PMML com variáveis resposta categóricas, o resultado do modelo é um conjunto de probabilidades preditas, uma para cada categoria, de o valor da resposta cair em cada categoria. A categoria com a maior probabilidade é considerada a categoria predita e utilizada na geração do gráfico de barras descrito para a configuração **Função de Densidade de Probabilidade** acima. Selecionar **Categoria predita** gera o gráfico de barras. Selecionar **Probabilidades preditas** gera histogramas da distribuição de probabilidades preditas para cada uma das categorias da resposta.

Agrupamento para análise de sensibilidade. Simulações que incluem análise de sensibilidade geram um conjunto independente de valores de resposta preditos para cada iteração definida pela análise (uma iteração para cada valor da entrada que está sendo variada). Quando as iterações estão presentes, o gráfico de barras da categoria predita para uma variável resposta categórica é exibido como um gráfico de barras agrupadas que inclui os resultados de todas as iterações. É possível optar por agrupar categorias ou agrupar iterações.

Saída

Gráficos Tornados. Os gráficos de tornados são gráficos de barras que exibem os relacionamentos entre as respostas e entradas simuladas usando uma variedade de métricas.

- **Correlação da resposta com a entrada.** Esta opção cria um gráfico de tornado dos coeficientes de correlação entre uma determinada resposta e cada uma de suas entradas simuladas. Esse tipo de gráfico de tornado não suporta respostas com um nível de medição nominal ou ordinal ou entradas simuladas com uma distribuição categórica.
- **Contribuição para variância.** Esta opção cria um gráfico de tornado que exibe a contribuição para a variância de uma resposta a partir de cada uma de suas entradas simuladas, permitindo avaliar até que grau cada entrada contribui para a incerteza geral na resposta. Esse tipo de gráfico de tornado não suporta respostas com níveis de medição ordinais ou nominais ou entradas simuladas com qualquer uma das seguintes distribuições: categórica, Bernoulli, binomial, Poisson ou binomial negativa.

- **Sensibilidade de resposta para alterar.** Esta opção cria um gráfico de tornado que exibe o efeito na resposta de modular cada entrada simulada por mais ou menos um número especificado de desvios padrão da distribuição associada à entrada. Esse tipo de gráfico de tornado não suporta respostas com níveis de medição ordinais ou nominais ou entradas simuladas com qualquer uma das seguintes distribuições: categórica, Bernoulli, binomial, Poisson ou binomial negativa.

Box plots de distribuições de resposta. Box plots estão disponíveis para variáveis resposta contínuas. Selecione **Sobrepor resultados a partir de respostas separadas** se o seu modelo preditivo possuir diversas variáveis resposta contínuas e você desejar exibir box plots para todas as respostas em um único gráfico.

Gráficos de dispersão de respostas versus entradas. Os gráficos de dispersão de respostas versus entradas simuladas estão disponíveis para variáveis resposta contínuas e categóricas e incluem gráficos de dispersão da resposta com entradas contínuas e categóricas. Os gráficos de dispersão que envolvem uma variável resposta categórica ou uma entrada categórica são exibidos como um heat map.

Criar uma tabela de valores de percentil. Para variáveis resposta contínuas, é possível obter uma tabela de percentis especificados das distribuições de resposta. Quartis (os 25⌂, 50⌂ e 75⌂ percentis) dividem as observações em quatro grupos de tamanhos iguais. Se desejar um número igual de grupos diferente de quatro, selecione **Intervalos** e especifique o número. Selecione **Percentis Customizados** para especificar percentis individuais - por exemplo, o 99⌂ percentil.

Estatísticas descritivas de distribuições de resposta. Esta opção cria tabelas de estatísticas descritivas para variáveis resposta contínuas e categóricas e também para entradas contínuas. Para variáveis resposta contínuas, a tabela inclui a média, o desvio padrão, a mediana, o mínimo e o máximo, o intervalo de confiança da média no nível especificado e os 5º e 95⌂ percentis da distribuição de resposta. Para variáveis resposta categóricas, a tabela inclui a porcentagem de casos que caem em cada categoria da resposta. Para variáveis resposta categóricas de modelos PMML, a tabela também inclui a probabilidade média de cada categoria da resposta. Para entradas contínuas, a tabela inclui a média, o desvio padrão e o mínimo e o máximo.

Correlações e tabela de contingência para entradas. Esta opção exibe uma tabela de coeficientes de correlação entre entradas simuladas. Quando entradas com distribuições categóricas são geradas a partir de uma tabela de contingência, a tabela de contingência dos dados que são gerados para essas entradas também é exibida.

Entradas simuladas para inclusão na saída. Por padrão, todas as entradas simuladas são incluídas na saída. É possível excluir as entradas simuladas selecionadas da saída. Isso excluirá os gráficos de tornado, de dispersão e saída tabular.

Limitar intervalos para variáveis resposta contínuas. É possível especificar o intervalo de valores válidos para uma ou mais variáveis resposta contínuas. Valores fora do intervalo especificado são excluídos de toda a saída e análises associadas às respostas. Para configurar um limite inferior, selecione **Inferior** na coluna Limite e insira um valor na coluna Mínimo. Para configurar um limite superior, selecione **Superior** na coluna Limite e insira um valor na coluna Máximo. Para configurar um limite inferior e um limite superior, selecione **Ambos** na coluna de Limite e insira os valores nas colunas Mínimo e Máximo.

Exibir Formatos. É possível configurar o formato utilizado ao exibir valores de resposta e de entradas (de entradas fixas e simuladas).

Salvar

Salvar o plano para esta simulação. É possível salvar as especificações atuais para sua simulação em um arquivo de plano de simulação. Os arquivos de plano de simulação possuem a extensão *.splan*. É possível reabrir o plano no Construtor de Simulação e, opcionalmente, fazer modificações e executar a simulação. É possível compartilhar o plano de simulação com outros usuários que, em seguida, poderão executá-lo no diálogo Executar Simulação. Os planos de simulação incluem todas as especificações, exceto as

seguintes: configurações para Funções de Densidade, configurações de saída para gráficos e tabelas, configurações de Opções Avançadas para Ajuste, Distribuição Empírica e Semente Aleatória.

Salvar os dados simulados como um novo arquivo de dados. É possível salvar as entradas simuladas, entradas fixas e valores de resposta preditos em um arquivo de dados do SPSS Statistics, em um novo conjunto de dados na sessão atual ou em um arquivo Excel. Cada caso (ou linha) do arquivo de dados consiste nos valores preditos das respostas junto das entradas simuladas e entradas fixas que geram os valores de resposta. Quando a análise de sensibilidade é especificada, cada iteração dá origem a um conjunto de casos contínuos que são rotulados com o número da iteração.

Executar diálogo de simulação

O diálogo Executar Simulação foi projetado para usuários que possuem um plano de simulação e desejam primariamente executar a simulação. Ele também fornece os recursos necessários para executar a simulação sob diferentes condições. Permite executar as seguintes tarefas gerais:

- Configurar ou modificar análise de sensibilidade para investigar o efeito de variar o valor de uma entrada fixa ou variar um parâmetro de distribuição para uma entrada simulada.
- Ajustar novamente as distribuições de probabilidade para entradas incertas (e correlações entre essas entradas) para novos dados.
- Modifique a distribuição para uma entrada simulada.
- Customizar saída.
- Execute a simulação.

Guia de Simulação

A guia Simulação permite especificar análise de sensibilidade, reajustar distribuições de probabilidade para as entradas simuladas e correlações entre entradas simuladas para os novos dados e modificar a distribuição de probabilidade associada a uma entrada simulada.

A grade de entradas Simuladas contém uma entrada para cada campo de entrada que estiver definido no plano de simulação. Cada entrada exibe o nome da entrada e o tipo de distribuição de probabilidade associado à entrada, junto de um gráfico de amostra da curva de distribuição associada. Cada entrada também possui um ícone de status associado (um círculo colorido com um visto) que é útil ao reajustar distribuições para novos dados. Além disso, as entradas podem incluir um ícone de bloqueio que indica que a entrada está bloqueada e não pode ser modificada ou reajustada para novos dados no diálogo Executar Simulação. Para modificar uma entrada bloqueada, será necessário abrir o plano de simulação no Construtor de Simulação.

Cada entrada será simulada ou fixa. Entradas simuladas são aquelas cujos valores são incertos e serão geradas ao desenhar a partir de uma distribuição de probabilidade especificada. Entradas fixas são aquelas cujos valores são conhecidos e que permanecem constantes para cada caso gerado na simulação. Para trabalhar com uma entrada específica, selecione a entrada na grade de entradas Simuladas.

Especificando análise de sensibilidade

A análise de sensibilidade permite investigar o efeito das mudanças sistemáticas em uma entrada fixa ou em um parâmetro de distribuição para uma entrada simulada ao gerar um conjunto independente de casos simulados – efetivamente, uma simulação separada - para cada valor especificado. Para especificar análise de sensibilidade, selecione uma entrada fixa ou simulada e clique em **Análise de Sensibilidade**. A análise de sensibilidade é limitada a uma entrada fixa única ou a um parâmetro de distribuição único para uma entrada simulada. Consulte o tópico “Análise de Sensibilidade” na página 186 para obter informações adicionais.

Reajustando distribuições para novos dados

Para reajustar automaticamente as distribuições de probabilidade para as entradas simuladas (e correlações entre entradas simuladas) para os dados no conjunto de dados ativos:

1. Verifique se cada uma das entradas de modelo é correspondida com o campo correto no conjunto de dados ativo. Cada entrada simulada é ajustada para o campo no conjunto de dados ativo especificado na lista suspensa **Campo** associada a essa entrada. É possível identificar facilmente entradas que não são correspondidas ao procurar entradas com um ícone de status que inclui um visto com um ponto de interrogação, conforme mostrado abaixo.

2. Modifique qualquer correspondência de campo necessária ao selecionar **Ajustar para um campo no conjunto de dados** e selecionar o campo na lista.
3. Clique em **Ajustar Tudo**.

Para cada entrada que tiver sido ajustada, a distribuição que mais próximo ajustar os dados será exibida junto de um gráfico da distribuição sobreposto em um histograma (ou gráfico de barras) dos dados históricos. Se um ajuste aceitável não puder ser localizado, então a distribuição Empírica será utilizada. Para entradas que se ajustam à distribuição Empírica, você verá somente um histograma dos dados históricos porque a distribuição Empírica é na realidade representada por esse histograma.

Nota: Para obter uma lista completa de ícones de status, consulte o tópico “Campos Simulados” na página 183.

Modificando distribuições de probabilidade

É possível modificar a distribuição de probabilidade para uma entrada simulada e, opcionalmente, alterar uma entrada simulada para uma entrada fixa ou vice-versa.

1. Selecione a entrada e selecione **Configurar manualmente a distribuição**.
2. Selecione o tipo de distribuição e especifique os parâmetros da distribuição. Para alterar uma entrada simulada para uma entrada fixa, selecione Fixo na lista suspensa **Tipo**.

Após ter inserido os parâmetros para uma distribuição, o gráfico de amostra da distribuição (exibido na entrada) será atualizado para refletir suas mudanças. Para obter mais informações sobre como especificar manualmente as distribuições de probabilidade, consulte o tópico “Campos Simulados” na página 183.

Incluir valores omissos de usuário de entradas categóricas ao ajustar. Isso especifica se valores omissos de usuário de entradas com uma Distribuição Categórica são tratados como válidos quando estiver reajustando os dados no conjunto de dados ativo. Valores omissos de sistema e valores omissos de usuário para todos os outros tipos de entradas são sempre tratados como inválidos. Todas as entradas devem ter valores válidos para um caso a ser incluído no ajuste de distribuição e no cálculo de correlações.

Guia de saída

A guia Saída permite customizar a saída gerada pela simulação.

Funções de Densidade. As funções de densidade são as médias primárias de análise do conjunto de resultados de sua simulação.

- **Função de Densidade de Probabilidade.** A função de densidade de probabilidade exibe a distribuição dos valores de resposta, permitindo determinar a probabilidade de a resposta estar em uma determinada região. Para respostas com um conjunto fixo de resultados - como "serviço insatisfatório", "serviço regular", "serviço bom" e "serviço excelente" - um gráfico de barras é gerado, que exibe a porcentagem de casos que caem em cada categoria da resposta.

- **Função de Distribuição Cumulativa.** A função de distribuição cumulativa exibe a probabilidade de que o valor da resposta seja menor ou igual a um valor especificado.

Gráficos de Tornado. Os gráficos de tornados são gráficos de barras que exibem os relacionamentos entre as respostas e entradas simuladas usando uma variedade de métricas.

- **Correlação da resposta com a entrada.** Esta opção cria um gráfico de tornado dos coeficientes de correlação entre uma determinada resposta e cada uma de suas entradas simuladas.
- **Contribuição para variância.** Esta opção cria um gráfico de tornado que exibe a contribuição para a variância de uma resposta a partir de cada uma de suas entradas simuladas, permitindo avaliar até que grau cada entrada contribui para a incerteza geral na resposta.
- **Sensibilidade de resposta para alterar.** Esta opção cria um gráfico de tornado que exibe o efeito na resposta de modular cada entrada simulada por mais ou menos um desvio padrão da distribuição associada à entrada.

Gráficos de dispersão de respostas versus entradas. Esta opção gera gráficos de dispersão de respostas versus entradas simuladas.

Box plots de distribuições de resposta. Esta opção gera box plots das distribuições de respostas.

Tabela de quartis. Esta opção gera uma tabela do quartis das distribuições de respostas. Os quartis de uma distribuição são 25⌂, 50⌂ e 75⌂ percentis da distribuição, dividindo as observações em quatro grupos de tamanhos iguais.

Correlações e tabela de contingência para entradas. Esta opção exibe uma tabela de coeficientes de correlação entre entradas simuladas. Uma tabela de contingência de associações entre entradas com uma distribuição categórica é exibida quando o plano de simulação especifica a geração de dados categóricos a partir de uma tabela de contingência.

Sobrepor resultados a partir de respostas separadas. Se o modelo preditivo que estiver simulando contiver diversas respostas, será possível especificar se os resultados das respostas separadas são exibidos em um único gráfico. Esta configuração se aplica a gráficos para funções de densidade de probabilidade, funções de distribuição cumulativa e box plots. Por exemplo, se essa opção for selecionada, as funções de densidade de probabilidade para todas as respostas serão exibidas em um único gráfico.

Salvar o plano para esta simulação. É possível salvar quaisquer modificações feitas em sua simulação em um arquivo de plano de simulação. Os arquivos de plano de simulação possuem a extensão *.splan*. É possível reabrir o plano no diálogo Executar Simulação ou no Construtor de Simulação. Os planos de simulação incluem todas as especificações, exceto configurações de saída.

Salvar os dados simulados como um novo arquivo de dados. É possível salvar as entradas simuladas, entradas fixas e valores de resposta preditos em um arquivo de dados do SPSS Statistics, em um novo conjunto de dados na sessão atual ou em um arquivo Excel. Cada caso (ou linha) do arquivo de dados consiste nos valores preditos das respostas junto das entradas simuladas e entradas fixas que geram os valores de resposta. Quando a análise de sensibilidade é especificada, cada iteração dá origem a um conjunto de casos contínuos que são rotulados com o número da iteração.

Se precisar de customização adicional de saída que está disponível aqui, considere executar sua simulação a partir do Construtor de Simulação. Consulte o tópico “Para executar uma simulação a partir de um plano de simulação” na página 179 para obter mais informações

Trabalhando com saída de gráfico a partir da simulação

Uma série dos gráficos gerados a partir de uma simulação têm recursos interativos que permitem customizar a exibição. Os recursos interativos estão disponíveis ao ativar (dando um clique duplo) o objeto de gráfico no Visualizador de Saída. Todos os gráficos de simulação são visualizações de gráfico.

Gráficos de função de densidade de probabilidade para variáveis resposta contínuas. Este gráfico possui duas linhas de referência verticais deslizantes que dividem o gráfico em regiões separadas. A tabela abaixo do gráfico exibe a probabilidade de que o destino está em cada uma das regiões. Se funções de densidade múltiplas são exibidas no mesmo gráfico, a tabela possui uma linha separada para as probabilidades associadas a cada função de densidade. Cada uma das linhas de referência possui uma régua de controle (triângulo invertido) que permite mover facilmente a linha. Uma série de recursos adicionais estão disponíveis clicando no botão **Opções de gráfico** no gráfico. Em particular, é possível configurar explicitamente as posições das réguas de controle, incluir linhas de referência fixas e alterar a visualização do gráfico a partir de uma curva contínua para um histograma ou vice-versa. Consulte o tópico "Opções de Gráfico" para obter mais informações

Gráficos de função de distribuição acumulativa para variáveis resposta contínuas. Este gráfico possui as duas mesmas linha de referência verticais que podem ser movidas e a tabela associada descrita para o gráfico de função de densidade de probabilidade acima. Ela também fornece acesso ao diálogo Opções de Gráfico, que permite configurar explicitamente as posições das réguas de controle, incluir linhas de referência fixas e especificar se a função de distribuição acumulativa é exibida como uma função crescente (o padrão) ou uma função decrescente. Consulte o tópico "Opções de Gráfico" para obter mais informações

Gráficos de barras para variáveis respostas categóricas com iteração de análise de sensibilidade. Para variáveis respostas categóricas com iterações de análise de sensibilidade, os resultados para a categoria de destino predita são exibidos como um gráfico de barras agrupadas que inclui os resultados de todas as iterações. O gráfico inclui uma lista suspensa que permite agrupar na categoria ou na iteração. Para modelos de cluster de duas etapas e modelos de cluster de K-Médias, é possível escolher agrupar em número de cluster ou iteração.

Box plots para respostas múltiplas com iterações de análise de sensibilidade. Para modelos preditivos com variáveis respostas contínuas e iterações de análise de sensibilidade, escolher exibir box plots para todos os destinos em um único gráfico produz um box plot em cluster. O gráfico inclui uma lista suspensa que permite agrupar na resposta ou na iteração.

Opções de Gráfico

O diálogo Opções de Gráfico permite customizar a exibição de gráficos ativados de funções de densidade de probabilidade e de funções de distribuição cumulativa gerados a partir de uma simulação.

Visualizar. A lista suspensa **Visualizar** se aplica apenas ao gráfico de função de densidade de probabilidade. Ele permite alternar a visualização do gráfico de uma curva contínua para um histograma. Este recurso não está disponível quando diversas funções de densidade forem exibidas no mesmo gráfico. Nesse caso, as funções de densidade podem ser visualizadas somente como curvas contínuas.

Ordem. A lista suspensa **Ordem** se aplica somente ao gráfico de função de distribuição cumulativa. Ela especifica se a função de distribuição cumulativa é exibida como uma função crescente (o padrão) ou como uma função decrescente. Quando exibida como uma função decrescente, o valor da função em um determinado ponto no eixo horizontal é a probabilidade de que a resposta esteja à direita desse ponto.

Posições da régua de controle. É possível configurar explicitamente as posições das linhas de referência deslizantes ao inserir valores nas caixas de texto Superior e Inferior. É possível remover a linha à esquerda ao selecionar **-Infinito**, configurando efetivamente a posição para infinito negativo, e remover a linha à direita ao selecionar **Infinito**, configurando efetivamente sua posição para infinito.

Linhas de referência. É possível incluir várias linhas de referência verticais fixas nas funções de densidade de probabilidade e nas funções de distribuição cumulativa. Quando diversas funções são exibidas em um único gráfico (devido a diversas respostas ou resultados de iterações da análise de sensibilidade), é possível especificar as funções específicas às quais as linhas serão aplicadas.

- **Sigmas.** É possível incluir linhas de referência em mais e menos um número especificado de desvios padrão da média de uma resposta.
- **Percentis.** É possível incluir linhas de referência em um ou dois valores de percentil da distribuição de uma resposta ao inserir valores nas caixas de texto Inferior e Superior. Por exemplo, um valor de 95 na caixa de texto Superior representa o 95⌂ percentil, que é o valor abaixo do qual 95% das observações caem. Da mesma forma, um valor de 5 na caixa de texto Inferior representa o 5⌂ percentil, que é o valor abaixo do qual 5% das observações caem.
- **Posições customizadas.** É possível incluir linhas de referência com valores especificados ao longo do eixo horizontal.

Rotular linhas de referência. Essa opção controla se os rótulos são aplicados às linhas de referência selecionadas.

As linhas de referência são removidas ao desmarcar a opção associada no diálogo Opções de Gráfico e clicar em **Continuar**.

Capítulo 35. Modelagem Geoespacial

As técnicas de modelagem geoespacial são projetadas para descobrir padrões nos dados que incluírem um componente geoespacial (mapa). O Assistente de Modelagem Geoespacial fornece métodos para analisar dados geoespaciais com e sem um componente de tempo.

Localizar associações com base em dados do evento e geoespaciais (Regras de Associação Geoespacial)
Usando as regras de associação geoespaciais, é possível localizar padrões nos dados com base em propriedades espaciais e não espaciais. Por exemplo, é possível identificar padrões nos dados criminais por localização e atributos demográficos. A partir desses padrões, é possível construir regras que preveem onde determinados tipos de crimes provavelmente ocorrerão.

Fazer previsões utilizando séries temporais e dados geoespaciais (Spatio-Temporal Prediction)
A predição temporal espacial utiliza dados que contêm dados de localização, campos de entrada para predição (preditores), um ou mais campos de tempo e um campo de destino. Cada localização possui várias linhas nos dados que representam os valores de cada preditor e o destino em cada intervalo de tempo.

Utilizando o Assistente de Modelagem Geoespacial

1. Nos menus, escolha:

   Analisar > **Modelagem Espacial e Temporal** > **Modelagem Espacial**
2. Siga as etapas do assistente.

Exemplos

Exemplos detalhados estão disponíveis no sistema de ajuda.

- Regras de associação geoespacial : **Ajuda** > **Tópicos** > **Estudos de Caso** > **Base de Estatísticas** > **Regras de associação espacial**
- Predição temporal espacial: **Ajuda** > **Tópicos** > **Estudos de Caso** > **Base de Estatísticas** > **Predição temporal espacial**

Selecionando Mapas

A modelagem geoespacial pode utilizar uma ou mais origens de dados do mapa. As origens de dados do mapa contêm informações que definem áreas geográficas e outros recursos geográficos, como estradas ou rios. Muitas origens do mapa também contêm dados demográficos ou outros dados descritivos e dados do evento, como relatórios criminais ou taxas de desemprego. É possível utilizar um arquivo de especificação de mapa definido anteriormente ou definir especificações de mapa aqui e salvar essas especificações para uso subsequente.

Carregar uma Especificação de Mapa
Carrega um arquivo de especificação de mapa (.mplan) definido anteriormente. As origens de dados do mapa que você define aqui podem ser salvas em um arquivo de especificação de mapa. Para predição temporal espacial, se você selecionar um arquivo de especificação de mapa que identifica mais de um mapa, será solicitado a selecionar um mapa no arquivo.

Incluir Arquivo de Mapa
Inclua um arquivo de formas ESRI (.shp) ou um archive .zip que contenha um arquivo de formas ESRI.

- Deverá haver um arquivo .dbf correspondente na mesma localização que o arquivo .shp, e esse arquivo deve ter o mesmo nome raiz que o arquivo .shp.
- Se o arquivo for um archive .zip, os arquivos .shp e .dbf deverão ter o mesmo nome raiz que o archive .zip.

- Se não houver nenhum arquivo de projeção (.prj) correspondente, será solicitado a selecionar um sistema de projeção.

Relacionamento
Para regras de associação geoespacial, essa coluna define como os eventos se relacionam aos recursos no mapa. Essa configuração não está disponível para predição temporal espacial.

Mover para Cima, Mover para baixo
A ordem da estrato dos elementos do mapa é determinada pela ordem em que eles aparecem na lista. O primeiro mapa na lista é a estrato inferior.

Selecionando um Mapa

Para predição temporal espacial, se você selecionar um arquivo de especificação de mapa que identifica mais de um mapa, será solicitado a selecionar um mapa no arquivo. A predição temporal espacial não suporta diversos mapas.

Relacionamento Geoespacial

Para regras de associação geoespacial, o diálogo Relacionamento Geoespacial define como os eventos se relacionam aos recursos no mapa.

- Essa configuração se aplica somente a regras de associação geoespaciais.
- Essa configuração afeta apenas as origens de dados associadas aos mapas que estiverem especificados como dados de contexto no passo para selecionar origens de dados.

Relacionamento

Fechar O evento ocorre próximo a um ponto ou área especificada no mapa.

Dentro de
O evento ocorre dentro de uma área especificada no mapa.

Contém
A área de evento contém um objeto de contexto de mapa.

Faz intersecção com
As localizações onde as linhas ou regiões de diferentes mapas se cruzam.

Cruzamento
Para diversos mapas, refere-se às localizações onde as linhas (para estradas, rios, ferrovias) a partir de linhas diferentes se cruzam.

Norte, Sul, Leste, Oeste
O evento ocorre dentro de uma área ao norte, sul, leste ou oeste de um ponto especificado no mapa.

Configurar Sistema de Coordenadas

Se não houver nenhum arquivo de projeção (.prj) com o mapa ou se você definir dois campos de uma origem de dados como um conjunto de coordenadas, deve-se configurar o sistema de coordenadas.

Geográfico padrão (longitude e latitude)
O sistema de coordenadas é longitude e latitude.

Cartesiano Simples (X e Y)
O sistema de coordenadas é coordenadas X e Y simples.

Usar um ID Bem Conhecidos (WKID)
"ID bem conhecido" para projeções comuns.

Use um Nome do Sistema de Coordenadas
O sistema de coordenadas baseia-se na projeção nomeada. O nome é colocado entre parênteses.

Configurando a Projeção

Se o sistema de projeção não puder ser determinado a partir das informações fornecidas com o mapa, será necessário especificar o sistema de projeção. A causa mais comum desta condição é a ausência de um arquivo de projeção (.prj) associado ao mapa ou um arquivo de projeção que não pode ser utilizado.

- **Uma cidade, região ou país (Mercator)**
- **Um país grande, vários países ou continentes (Winkel Tripel)**
- **Uma área muito próxima ao equador (Mercator)**
- **Uma área próxima a um dos polos (Estereográfico)**

A projeção Mercator é uma projeção comum utilizada em muitos mapas. Essa projeção trata o globo como um cilindro que é desenrolado em uma superfície plana. A projeção Mercator distorce o tamanho e a forma de objetos grandes. Esta distorção aumenta conforme você move para longe do equador e se aproxima dos polos. As projeções Winkel Tripel e Estereográfico fazem ajustes para o fato de que um mapa representa uma parte de uma esfera tridimensional exibida em duas dimensões.

Sistema de Projeção e de Coordenadas

Se você selecionar mais de um mapa e os mapas possuírem sistemas de projeção e de coordenadas diferentes, o mapa deverá ser selecionado com o sistema de projeção que deseja utilizar. Esse sistema de projeção será usado para todos os mapas quando eles forem combinados em uma saída.

Origens de Dados

Uma origem de dados pode ser um arquivo dBase que é fornecido com o arquivo de formas, com um arquivo de dados do IBM SPSS Statistics ou com um conjunto de dados aberto na sessão atual.

Dados de Contexto. Os dados de contexto identificam recursos no mapa. Os dados de contexto também podem conter campos que podem ser utilizados como entradas para o modelo. Para utilizar um arquivo dBase de contexto (.dbf) que está associado a um arquivo de formas de mapa (.shp), o arquivo dBase de contexto deve estar na mesma localização que o arquivo de formas e deve ter o mesmo nome raiz. Por exemplo, se o arquivo de formas for `geodata.shp`, o arquivo dBase deverá ser denominado `geodata.dbf`

Dados de Eventos. Os dados de eventos contêm informações sobre eventos que ocorrem, como crimes ou acidentes. Esta opção está disponível apenas para regras de associação geoespaciais.

Densidade de Ponto. Intervalo de tempo e dados de coordenadas para estimativas da densidade kernel. Esta opção está disponível apenas para predição temporal espacial.

Incluir. Abre um diálogo para incluir origens de dados. Uma origem de dados pode ser um arquivo dBase que é fornecido com o arquivo de formas, com um arquivo de dados do IBM SPSS Statistics ou com um conjunto de dados aberto na sessão atual.

Associar. Abre um diálogo para especificar os identificadores (coordenadas ou chaves) utilizados para associar dados aos mapas. Cada origem de dados deve conter um ou mais identificadores que associam os dados ao mapa. Arquivos dBase provenientes de um arquivo de formas geralmente contêm um campo que é usado automaticamente como o identificador de padrão. Para outras origens de dados, deve-se especificar os campos que são utilizados como identificadores.

Validar Chave. Abre um diálogo para validar correspondência de chave entre o mapa e a origem de dados.

Regras de associação geoespacial

- Pelo menos uma origem de dados deve ser uma origem de dados do evento.

- Todas as origens de dados do evento devem utilizar a mesma forma de identificadores de associação do mapa: coordenadas ou valores de chave.
- Se as origens de dados do evento forem associadas aos mapas com valores de chave, todas as origens de eventos deverão usar o mesmo tipo de recurso de mapa (por exemplo, polígonos, pontos, linhas).

Predição temporal espacial

- Deve haver uma origem de dados de contexto.
- Se houver somente uma origem de dados (um arquivo de dados sem nenhum mapa associado), ele deverá incluir valores de coordenadas.
- Se você tiver duas origens de dados, uma origem de dados deverá ser dados de contexto, e a outra origem de dados deverá ser dados de densidade de ponto.
- Não é possível incluir mais de duas origens de dados.

Incluir uma Origem de Dados

Uma origem de dados pode ser um arquivo dBase que é fornecido com o arquivo de formas ou arquivo de contexto, com um arquivo de dados do IBM SPSS Statistics ou com um conjunto de dados aberto na sessão atual.

É possível incluir a mesma origem de dados diversas vezes se desejar usar uma associação espacial diferente em cada uma das vezes.

Associação de Dados e Mapa

Cada origem de dados deve conter um ou mais identificadores que associam os dados ao mapa.

Coordenadas
: A origem de dados contém campos que representam coordenadas Cartesianas, portanto, selecione os campos que representam essas coordenadas X e Y. Para regras de associação geoespaciais, também pode haver uma coordenada Z.

Valores da chave
: Os valores da chave nos campos na origem de dados correspondem às chaves de mapa selecionadas. Por exemplo, um mapa de regiões pode ter um identificador de nome (chave de mapa) rotulando cada região. Esse identificador corresponde a um campo nos dados que também contém os nomes das regiões (chave de dados). Os campos são correspondidos com as chaves de mapa com base na ordem em que eles são exibidos nas duas listas.

Validar chaves

O diálogo Validar Chaves fornece uma sumarização da correspondência de registro entre o mapa e a origem de dados, com base nas chaves do identificador selecionado. Se houver valores de chave de dados não correspondentes, será possível correspondê-los manualmente aos valores de chave de mapa.

Regras de associação geoespacial

Para regras de associação geoespaciais, depois de definir mapas e origens de dados, as etapas restantes no assistente são:

- Se houver diversas origens de dados do evento, defina como as origens de dados do evento são mescladas.
- Selecione os campos a serem usados como condições e previsões na análise.

Opcionalmente, também é possível:

- Selecionar opções de saída diferentes.
- Salvar um arquivo de modelo de pontuação.
- Criar novos campos para valores e regras preditos nas origens de dados usadas no modelo.

- Customizar configurações para construção de regras de associação.
- Customizar configurações de categorização e agregação.

Definir Campos de Dados do Evento

Para regras de associação geoespacial, se houver mais de uma origem de dados do evento, as origens de dados do evento serão mescladas.

- Por padrão, apenas os campos que forem comuns a todas as origens de dados do evento são incluídos.
- É possível exibir uma lista de campos comuns, campos para uma origem de dados específica ou campos a partir de todas as origens de dados e selecionar os campos que você deseja incluir.
- Para campos comuns, o **Tipo** e a **Medição** devem ser os mesmos para todas as origens de dados. Se houver conflitos, será possível especificar o tipo e o nível de medição para uso em cada campo comum.

Selecionar Campos

A lista de campos disponíveis inclui campos das origens de dados do evento e campos das origens de dados de contexto.

- É possível controlar a lista de campos exibidos ao selecionar uma origem de dados a partir da lista **Origens de Dados**.
- Deve-se selecionar pelo menos dois campos. Pelo menos um deve ser uma condição, e pelo menos um deve ser uma predição. Existem diversas formas para atender a esse requisito, incluindo selecionar dois campos para a lista **Ambos (Condição e Predição)**.
- As regras de associação preveem valores dos campos de predição que são baseados nos valores dos campos condição. Por exemplo, na regra "If x=1 e y=2, then z=3", os valores de x e y são condições, e o valor de z é a predição.

Saída

Tabelas de Regras
Cada tabela de regras exibe as principais regras e valores de confiança, suporte de regra, elevação, suporte de condição e implementabilidade. Cada tabela é armazenada pelos valores dos critérios selecionados. É possível exibir todas as regras ou o **Número** de regras principais, com base no critério selecionado.

Nuvem de palavra classificável
Uma lista de regras principais com base nos valores do critério selecionado. O tamanho do texto indica a importância relativa da regra. O objeto de resultado interativo contém as regras principais para confiança, suporte de regra, elevação, suporte de condição e implementabilidade. O critério selecionado determina qual lista de regras é exibida por padrão. É possível selecionar um critério diferente de forma interativa na saída. **Máximo de regras para exibir** determina o número de regras que são exibidas na saída.

Mapa Gráfico de barras e mapa interativos das regras principais, com base no critério selecionado. Cada objeto de resultado interativo contém as regras principais para confiança, suporte de regra, elevação, suporte de condição e implementabilidade. O critério selecionado determina qual lista de regras é exibida por padrão. É possível selecionar um critério diferente de forma interativa na saída. **Máximo de regras para exibir** determina o número de regras que são exibidas na saída.

Tabelas de informações de modelo

Transformações de Campo.
Descreve as transformações que são aplicadas aos campos utilizados na análise.

Sumarização do Registro.
O número e a porcentagem de registros incluídos e excluídos.

Estatísticas de Regra.
Estatísticas básicas do suporte de condição, confiança, suporte de regra, elevação e implementabilidade. As estatísticas incluem média, o mínimo, o máximo e o desvio padrão.

Itens Mais Frequentes.
Itens que ocorrem com mais frequência. Um item é incluído em uma condição ou uma predição em uma regra. Por exemplo, idade < 18 ou gender=female.

Campos Mais Frequentes.
Campos que ocorrem com mais frequência nas regras.

Entradas Excluídas.
Campos que são excluídos da análise e o motivo pelo qual cada campo foi excluído.

Critério de Tabelas de Regras, Nuvem de Palavras e Mapas

Confiança.
A porcentagem de predições de regra correta.

Suporte de Regra.
A porcentagem de casos para os quais a regra é verdadeira. Por exemplo, se a regra for "If x=1 and y=2, then z=3," o suporte de regra será a porcentagem real de casos nos dados para os quais x=1, y=2 e z=3.

Elevação.
Elevação é uma medida do quanto a regra melhora a predição em comparação com a chance aleatória. É a razão de predições corretas para a ocorrência geral do valor predito. O valor deve ser maior que 1. Por exemplo, se o valor predito ocorrer 20% do tempo e a confiança na predição for 80%, então o valor de elevação será 4.

Suporte de Condição.
A porcentagem de casos para os quais a condição da regra existe. Por exemplo, se a regra for "If x=1 and y=2, then z=3", o suporte de condição é a proporção de casos nos dados para os quais x=1 e y=2.

Implementabilidade.
A porcentagem de predições incorretas quando as condições são verdadeiras. A implementabilidade é igual a (1-confiança) multiplicado pelo suporte de condição ou ao suporte de condição menos o suporte de regra.

Salve

Salvar o mapa e dados de contexto como uma especificação de mapa
Salve as especificações de mapa em um arquivo externo (.mplan). É possível carregar esse arquivo de especificação de mapa no assistente para análises subsequentes. Também é possível utilizar o arquivo de especificação de mapa com o comando `SPATIAL ASSOCIATION RULES`.

Copiar quaisquer arquivos de mapas e de dados na especificação
Dados de arquivos de formas de mapa, arquivos de dados externos e conjuntos de dados utilizados na especificação de mapa são salvos no arquivo de especificação de mapa.

Escoragem
Salva os melhores valores de regras, valores de confiança para as regras e os valores de ID numérico para as regras como novos campos na origem de dados especificada.

Origem de Dados para Escorar
Uma ou mais origens de dados em que os novos campos são criados. Se a origem de dados não estiver aberta na sessão atual, ela será aberta na sessão atual. O arquivo modificado deve ser salvo explicitamente para salvar os novos campos.

Valores de Resposta

Crie novos campos para os campos de destino (predição) selecionados.

- Dois novos campos são criados para cada campo de destino: valor predito e valor de confiança.
- Para campos de variáveis resposta contínuas (escala), o valor predito é uma sequência de caracteres que descreve um intervalo de valores. Um valor no formato "(value1, value2]" significa "maior que value1 e menor ou igual a value2".

Número de melhores regras

Crie novos campos para o número de melhores regras especificadas. Três novos campos são criados para cada regra: valor da regra, valor de confiança e um valor de ID numérico para a regra.

Prefixo do Nome

Prefixo a ser utilizado para os novos nomes de campo.

Construção de Regra

Os parâmetros de construção de regra configuram os critérios para as regras de associação geradas.

Itens por Regra

Número de valores de campo que podem ser incluídos nas condições da regra e predições. O número total de itens não pode exceder 10. Por exemplo, na regra "If x=1 and y=2, then z=3", há dois itens de condição e um item de predição.

Máximo de predições.

Número máximo de valores de campo que podem ocorrer nas predições para uma regra.

Máximo de condições.

Número máximo de valores de campo que podem ocorrer nas condições para uma regra.

Excluir par

Exclui os pares de campos especificados de serem incluídos na mesma regra.

Critérios de Regra

Confiança.

Confiança mínima que uma regra deve ter para ser incluída na saída. A confiança é a porcentagem de predições corretas.

Suporte de Regra.

Suporte de regra mínimo que uma regra deve ter para ser incluída na saída. O valor representa a porcentagem de casos para os quais a regra é verdadeira nos dados observados. Por exemplo, se a regra for "If x=1 and y=2, then z=3," o suporte de regra será a porcentagem real de casos nos dados para os quais x=1, y=2 e z=3.

Suporte de Condição.

Suporte de condição mínimo que uma regra deve ter para ser incluída na saída. O valor representa a porcentagem de casos para os quais a condição existe. Por exemplo, se a regra for "If x=1 and y=2, then z=3", o suporte de condição é a porcentagem de casos nos dados para os quais x=1 e y=2.

Elevação.

Elevação mínima que uma regra deve ter para ser incluída na saída. Elevação é uma medida do quanto a regra melhora a predição em uma chance aleatória. É a razão de predições corretas para a ocorrência geral do valor predito. Por exemplo, se o valor predito ocorrer 20% do tempo e a confiança na predição for 80%, então o valor de elevação será 4.

Tratar como iguais

Identifica pares de campos que devem ser tratados como o mesmo campo.

Categorização e Agregação

- A agregação é necessária quando houver mais registros nos dados do que recursos no mapa. Por exemplo, você possui registros de dados para estados individuais, mas também possui um mapa de estados.
- É possível especificar o método de medida de sumarização agregada para campos contínuos e ordinais. Os campos nominais são agregados com base no valor modal.

Contínuo
Para campos contínuos (escala), a medida de sumarização pode ser média, mediana ou soma.

Ordinal
Para os campos ordinais, a medida de sumarização pode ser mediana, modo, mais alta ou mais baixa.

Número de categorias
Configura o número máximo de categorias para campos contínuos (escala). Os campos contínuos são sempre agrupados ou "categorizados" em intervalos de valores. Por exemplo: menor ou igual a 5, maior que 5 e menor ou igual a 10 ou maior que 10.

Agregar o mapa
Aplique agregação para dados e mapas.

Configurações customizadas para campos específicos
É possível substituir a medida de sumarização padrão e o número de categorias para campos específicos.

- Clique no ícone para abrir o diálogo **Seletor de Campos** e selecione um campo para incluir na lista.
- Na coluna **Agregação**, selecione uma medida de sumarização.
- Para campos contínuos, clique no botão na coluna **Categorias** para especificar um número customizado de categorias para o campo no diálogo **Categorias**.

Predição temporal espacial

Para predição temporal espacial, após definir mapas e origens de dados, as etapas restantes no assistente são:

- Especificar o campo de destino, os campos de tempo e os preditores opcionais.
- Definir intervalos de tempo ou períodos cíclicos para campos de tempo.

Opcionalmente, também é possível:

- Selecionar opções de saída diferentes.
- Customizar parâmetros de construção de modelo.
- Customizar configurações de agregação.
- Salvar valores preditos em um conjunto de dados na sessão atual ou em um arquivo de dados formatados do IBM SPSS Statistics.

Selecionar Campos

A lista de campos disponíveis inclui campos de origens de dados selecionadas. É possível controlar a lista de campos exibidos ao selecionar uma origem de dados a partir da lista **Origens de Dados**.

Destino
Um campo de destino é necessário. O destino é o campo para o qual os valores são preditos.

- O campo de destino deve ser um campo numérico contínuo (escala).
- Se houver duas origens de dados, o destino é estimativas da densidade de kernel, e "Densidade" é exibido como o nome de destino. Essa seleção não pode ser alterada.

Preditores

Um ou mais campos preditores podem ser especificados. Esta configuração é opcional.

Campos de Tempo

Deve-se selecionar um ou mais campos que representem períodos de tempo ou selecionar **Períodos Cíclicos**.

- Se houver duas origens de dados, deve-se selecionar os campos de tempo a partir de ambas as origens de dados. Ambos os campos de tempo devem representar o mesmo intervalo.
- Para períodos cíclicos, deve-se especificar os campos que definem ciclos de periodicidade no painel Intervalos de Tempo do assistente.

Intervalos de Tempo

As opções nesse painel baseiam-se na opção de **Campos de Tempo** ou **Período cíclico** no passo para selecionar campos.

Campos de Hora

Campos de Tempo Selecionados. Se selecionar um ou mais campos de tempo no passo de seleção de campos, esses campos serão exibidos nesta lista.

Intervalo de Tempo. Selecione o intervalo de tempo apropriado na lista. Dependendo do intervalo de tempo, também é possível especificar outras configurações, como o intervalo entre as observações (incremento) e valores iniciais. Esse intervalo de tempo é utilizado para todos os campos de tempo selecionados.

- O procedimento supõe que todos os casos (registros) representam intervalos igualmente espaçados.
- Com base no intervalo de tempo selecionado, o procedimento poderá detectar observações omissas ou diversas observações no mesmo intervalo de tempo que precisam ser agregadas. Por exemplo, se o intervalo de tempo for dias e a data 2014-10-27 for seguida por 2014-10-29, então há uma observação omissa para 2014-10-28. Se o intervalo de tempo for mês, então diversas datas no mesmo mês serão agregadas.
- Para alguns intervalos de tempo, uma configuração adicional pode definir quebras nos intervalos normais igualmente espaçados. Por exemplo, se o intervalo de tempo for dias, mas apenas dias da semana forem válidos, será possível especificar que há cinco dias em uma semana, com a semana iniciando na segunda-feira.
- Se o campo de tempo selecionado não for um campo de formato de data ou de formato de hora, o intervalo de tempo será configurado automaticamente para **Períodos** e não poderá ser alterado.

Campos de ciclo

Se selecionar **Período cíclico** no passo para selecionar campos, deve-se especificar os campos que definem os períodos cíclicos. Um período cíclico identifica a variação cíclica repetitiva, como o número de meses em um ano ou o número de dias em uma semana.

- É possível especificar até três campos que definem períodos cíclicos.
- O primeiro campo de ciclo representa o nível mais alto do ciclo. Por exemplo, se houver uma variação cíclica por ano, trimestre e mês, o campo que representa o ano é o primeiro campo de ciclo.
- A duração do ciclo para os primeiro e segundo campos de ciclo é a periodicidade no nível subsequente. Por exemplo, se os campos de ciclo forem ano, trimestre e mês, a primeira duração do ciclo será 4 e a segunda duração do ciclo será 3.
- O valor de início para os segundos e terceiro campos de ciclo é o primeiro valor em cada um desses períodos cíclicos.
- Os valores de duração de ciclo e inicial devem ser números inteiros positivos.

Agregação

- Se selecionar quaisquer **Preditores** no passo para selecionar campos, será possível selecionar o método de sumarização de agregação para os preditores.
- A agregação é necessária quando houver mais de um registro em um intervalo de tempo definido. Por exemplo, se o intervalo de tempo for mês, então diversas datas no mesmo mês serão agregadas.
- É possível especificar o método de medida de sumarização de agregação para campos contínuos e ordinais. Os campos nominais são agregados com base no valor modal.

Contínuo
Para campos contínuos (escala), a medida de sumarização pode ser média, mediana ou soma.

Ordinal
Para os campos ordinais, a medida de sumarização pode ser mediana, modo, mais alta ou mais baixa.

Configurações customizadas para campos específicos
É possível substituir a medida de sumarização de agregação padrão para preditores específicos.

- Clique no ícone para abrir o diálogo **Seletor de Campos** e selecione um campo para incluir na lista.
- Na coluna **Agregação**, selecione uma medida de sumarização.

Saída

Mapas

Valores de destino.
Mapa de valores para o campo de destino selecionado.

Correlação
Mapa de correlações.

Agrupamentos
Mapa que destaca os clusters de localizações que são semelhantes entre si. Mapas de clusters estão disponíveis apenas para os modelos empíricos.

Limite de similaridade de localização
A similaridade que é necessária para criar os clusters. O valor deve ser um número maior que zero e menor que 1.

Especifique o número máximo de clusters.
O número máximo de clusters para exibir.

Tabelas de Avaliação de Modelo

Especificações de Modelo.
Sumarização das especificações que são utilizadas para executar a análise, incluindo campos de destino, de entrada e de localizações.

Sumarização de Informações Temporais.
Identifica os campos e intervalos de tempo que são utilizados no modelo.

Testes de Efeitos na Estrutura de Média.
A saída inclui valor de estatísticas do teste, graus de liberdade e o nível de significância para o modelo e cada efeito.

Estrutura de Média de Coeficientes do Modelo.
A saída inclui o valor do coeficiente, o erro padrão, o valor de estatísticas de teste, o nível de significância e os intervalos de confiança para cada termo modelo.

Coeficientes Autorregressivos.
A saída inclui o valor do coeficiente, o erro padrão, o valor de estatísticas de teste, o nível de significância e os intervalos de confiança para cada lag.

Testes de Covariância Espacial.
Para modelos paramétricos baseados em variograma, exibe os resultados de teste de qualidade de ajuste para a estruturas de covariâncias espaciais. Os resultados do teste podem determinar se é necessário modelar a estruturas de covariâncias espaciais parametricamente ou utilizar um modelo não paramétrico.

Covariância Espacial Paramétrica.
Para modelos paramétricos baseados em variograma, exibe as estimativas paramétrica para covariância espacial paramétrica.

Opções de Modelo

Configurações de modelo

Incluir automaticamente um intercepto
Inclui o intercepto no modelo.

Lag máximo de autorregressão
O lag máximo de autorregressão. O valor deve ser um número inteiro entre 1 e 5.

Covariância espacial
Especifica o método de estimação para covariância espacial.

Paramétrico
O método de estimação é paramétrico. O método pode ser **Gaussiano**, **Exponencial** ou **Potência Exponencial**. Para Potência Exponencial, é possível especificar o valor de **Potência**.

Não paramétrico
O método de estimação é não paramétrico.

Salvar

Salvar o mapa e dados de contexto como uma especificação de mapa
Salve as especificações de mapa em um arquivo externo (.mplan). É possível carregar esse arquivo de especificação de mapa no assistente para análises subsequentes. Também é possível utilizar o arquivo de especificação de mapa com o comando `SPATIAL TEMPORAL PREDICTION`.

Copiar quaisquer arquivos de mapas e de dados na especificação
Dados de arquivos de formas de mapa, arquivos de dados externos e conjuntos de dados que são utilizados na especificação de mapa são salvos no arquivo de especificação de mapa.

Escoragem
Salva valores preditos, a variância e limites de confiança superior e inferior para o campo de destino no arquivo de dados selecionado.

- É possível salvar valores preditos em um conjunto de dados aberto na sessão atual ou em um arquivo de dados de formato do IBM SPSS Statistics.
- O arquivo de dados não pode ser uma origem de dados que é utilizada no modelo.
- O arquivo de dados deve conter todos os campos de tempo e preditores que são utilizados no modelo.
- Os valores de tempo devem ser maiores que os valores de tempo utilizados no modelo.

Avançado

Máximo de casos com valores omissos (%)
A porcentagem máxima de casos com valores omissos.

Nível de significância
O nível de significância para determinar se um modelo paramétrico baseado em variograma é

apropriado. O valor deve ser maior que 0 e menor que 1. O valor padrão é 0,05. O nível de significância é utilizado no teste de Qualidade do ajuste para a estruturas de covariâncias espacial. A estatística de Qualidade do ajuste é utilizada para determinar se um modelo paramétrico ou não paramétrico deve ser utilizado.

Fator de incerteza (%)
O fator de incerteza é um valor percentual que representa o crescimento da incerteza para futuras previsões. Os limites superior e inferior de incerteza da previsão aumentam de acordo com a porcentagem especificada para cada passo no futuro.

Conclusão

No último passo do Assistente de Modelagem Geoespacial, é possível executar o modelo ou colar a sintaxe de comando gerada em uma janela de sintaxe. É possível modificar e salvar a sintaxe gerada para uso subsequente.

Avisos

Essas informações foram desenvolvidas para produtos e serviços oferecidos nos Estados Unidos. Esse material pode estar disponível a partir da IBM em outros idiomas. No entanto, pode ser necessário possuir uma cópia do produto ou da versão do produto nesse idioma para acessá-lo.

É possível que a IBM não ofereça produtos, serviços ou recursos discutidos neste documento em outros países. Consulte um representante IBM local para obter informações sobre produtos e serviços disponíveis atualmente em sua área. Qualquer referência a produtos, programas ou serviços IBM não significa que apenas produtos, programas ou serviços IBM possam ser utilizados. Qualquer produto, programa ou serviço funcionalmente equivalente, que não infrinja nenhum direito de propriedade intelectual da IBM poderá ser utilizado em substituição a este produto, programa ou serviço. Entretanto, a avaliação e verificação da operação de qualquer produto, programa ou serviço não IBM são de responsabilidade do Cliente.

A IBM pode ter patentes ou solicitações de patentes pendentes relativas a assuntos tratados nesta publicação. O fornecimento desta publicação não lhe garante direito algum sobre tais patentes. Pedidos de licença podem ser enviados, por escrito, para:

Gerência de Relações Comerciais e Industriais da IBM Brasil
Av. Pasteur, 138-146
CEP 22290-240
Rio de Janeiro, RJ
Brasil

Para pedidos de licença relacionados a informações de DBCS (Conjunto de Caracteres de Byte Duplo), entre em contato com o Departamento de Propriedade Intelectual da IBM em seu país ou envie pedidos de licença, por escrito, para:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

A INTERNATIONAL BUSINESS MACHINES CORPORATION FORNECE ESTA PUBLICAÇÃO "NO ESTADO EM QUE SE ENCONTRA", SEM GARANTIA DE NENHUM TIPO, SEJA EXPRESSA OU IMPLÍCITA, INCLUINDO, MAS NÃO SE LIMITANDO ÀS GARANTIAS IMPLÍCITAS DE NÃO-VIOLAÇÃO, COMERCIALIZAÇÃO OU ADEQUAÇÃO A UM DETERMINADO PROPÓSITO. Alguns países não permitem a exclusão de garantias explícitas ou implícitas em certas transações; portanto, esta instrução pode não se aplicar ao Cliente.

Essas informações podem conter imprecisões técnicas ou erros tipográficos. São feitas alterações periódicas nas informações aqui contidas; tais alterações serão incorporadas em futuras edições desta publicação. A IBM pode, a qualquer momento, aperfeiçoar e/ou alterar o(s) produto(s) e/ou programa(s) descritos nesta publicação, sem aviso prévio.

Qualquer referência nestas informações a websites não IBM são fornecidas apenas por conveniência e não representam de forma alguma um endosso a esses websites. Os materiais contidos nesses websites não fazem parte dos materiais para esse produto IBM e o uso desses websites é de inteira responsabilidade do Cliente.

A IBM por usar ou distribuir as informações fornecidas da forma que julgar apropriada sem incorrer em qualquer obrigação para com o Cliente.

Licenciados deste programa que desejam obter informações sobre o mesmo com o objetivo de permitir: (i) a troca de informações entre programas criados independentemente e outros programas (incluindo este) e (ii) o uso mútuo de informações trocadas, devem entrar em contato com:

Gerência de Relações Comerciais e Industriais da IBM Brasil
Av. Pasteur, 138-146
CEP 22290-240
Rio de Janeiro, RJ
Brasil

Tais informações podem estar disponíveis, sujeitas a termos e condições apropriadas, incluindo em alguns casos o pagamento de uma taxa.

O programa licenciado descrito nesta publicação e todo o material licenciado disponível são fornecidos pela IBM sob os termos do Contrato com o Cliente IBM, do Contrato Internacional de Licença do Programa IBM ou de qualquer outro contrato equivalente.

Os exemplos de dados de desempenho e do Cliente citados são apresentados apenas para propósitos ilustrativos. Resultados de desempenho reais podem variar dependendo das configurações específicas e das condições operacionais.

Informações relativas a produtos não IBM foram obtidas junto aos fornecedores dos respectivos produtos, de seus anúncios publicados ou de outras fontes disponíveis publicamente. A IBM não testou esses produtos e não pode confirmar a precisão de desempenho, compatibilidade nem qualquer outra reivindicação relacionada a produtos não IBM. Perguntas sobre os recursos de produtos não IBM devem ser endereçadas aos fornecedores desses produtos.

Instruções relativas à direção futura ou intento da IBM estão sujeitas a mudança ou retirada sem aviso e representam metas e objetivos apenas.

Estas informações contêm exemplos de dados e relatórios utilizados nas operações diárias de negócios. Para ilustrá-los da forma mais completa possível, os exemplos podem incluir nomes de assuntos, empresas, marcas e produtos. Todos esses nomes são fictícios e qualquer semelhança com pessoas ou empresas reais é mera coincidência.

LICENÇA DE COPYRIGHT:

Estas informações contêm programas de aplicativos de amostra na linguagem fonte, ilustrando as técnicas de programação em diversas plataformas operacionais. O Cliente pode copiar, modificar e distribuir estes programas de amostra sem a necessidade de pagar à IBM, com objetivos de desenvolvimento, utilização, marketing ou distribuição de programas aplicativos em conformidade com a interface de programação de aplicativo para a plataforma operacional para a qual os programas de amostra são criados. Esses exemplos não foram testados completamente em todas as condições. Portanto, a IBM não pode garantir ou implicar a confiabilidade, manutenção ou função destes programas. Os programas de amostra são fornecidos "NO ESTADO EM QUE SE ENCONTRAM", sem garantia de qualquer tipo. A IBM não será responsabilizada por quaisquer danos decorrentes do uso dos programas de amostra.

Cada cópia ou parte destes programas de amostra ou qualquer trabalho derivado deve incluir um aviso de copyright com os dizeres:

© nome de sua empresa) (ano). Partes deste código são derivadas dos Programas de Amostra da IBM Corp.

Marcas comerciais

IBM, o logotipo IBM e ibm.com são marcas comerciais ou marcas registradas da International Business Machines Corp., registradas em muitos países no mundo todo. Outros nomes de produtos e serviços podem ser marcas comerciais da IBM ou de outras empresas. ma lista atual de marcas comerciais da IBM está disponível na web em "Copyright and trademark information" em www.ibm.com/legal/copytrade.shtml.

Adobe, o logotipo Adobe, PostScript e o logotipo PostScript são marcas registradas ou marcas comerciais da Adobe Systems Incorporated nos Estados Unidos e/ou em outros países.

Intel, o logotipo Intel, Intel Inside, o logotipo Intel Inside, Intel Centrino, o logotipo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium são marcas comerciais ou marcas registradas da Intel Corporation ou de suas subsidiárias nos Estados Unidos e em outros países.

Linux é uma marca registrada da Linus Torvalds nos Estados Unidos, e/ou em outros países.

Microsoft, Windows, Windows NT e o logotipo Windows são marcas comerciais da Microsoft Corporation nos Estados Unidos e/ou em outros países.

UNIX é uma marca registrada da The Open Group nos Estados Unidos e em outros países.

Java e todas as marcas comerciais e logotipos baseados em Java são marcas comerciais ou marcas registradas da Oracle e/ou suas afiliadas.

Índice Remissivo

Numéricos

2 T de Hotelling
  na Análise de confiabilidade 165, 166

A

ajuste de distribuição
  na simulação 183
ajuste de distribuição automática
  na simulação 183
alocação de memória
  na Análise de Cluster TwoStep 110
alpha de Cronbach
  na Análise de confiabilidade 165, 166
amostra de treinamento
  na Análise do vizinho mais próximo 90
amostra de validação
  na Análise do vizinho mais próximo 90
amostras relacionadas 147, 150
amplitude múltipla de Ryan-Einot-Gabriel-Welsch
  no GLM 48
  no One-Way ANOVA 40
análise de cluster
  Análise de cluster por K-médias 123
  Análise de clusters hierárquica 119
  eficiência 124
Análise de cluster de duas etapas 109
  estatísticas 111
  opções 110
  salvar em arquivo externo 111
  salvar no arquivo de trabalho 111
Análise de cluster por K-médias
  associação de cluster 124
  critérios de convergência 124
  distâncias de cluster 124
  eficiência 124
  estatísticas 123, 125
  exemplos 123
  iterações 124
  métodos 123
  recursos adicionais do comando 125
  salvando informações do cluster 124
  valores omissos 125
  visão geral 123
Análise de clusters hierárquica 119
  armazenando casos em cluster 119
  armazenando variáveis em cluster 119
  associação de cluster 120
  dendrogramas 120
  estatísticas 119, 120
  example 119
  gráfico icicle 120
  matrizes de distância 120
  medidas de distância 119
  medidas de similaridade 119
  medidas de transformação 119
Análise de clusters hierárquica *(continuação)*
  métodos de clusterização 119
  orientação do gráfico 120
  planejamentos de aglomeração 120
  recursos adicionais do comando 120
  salvando novas variáveis 120
  valores de transformação 119
análise de componentes principais 103, 104
Análise de confiabilidade 165
  2 T de Hotelling 166
  coeficiente de correlação intraclasse 166
  correlações e covariâncias entre itens 166
  descritiva 166
  estatísticas 165, 166
  example 165
  Kuder-Richardson 20 166
  recursos adicionais do comando 167
  Tabela de ANOVA 166
  Teste de aditividade de Tukey 166
análise de múltiplas respostas
  Frequências de múltiplas respostas 154
  tabelas de frequências 154
  tabulação cruzada 155
  Tabulações cruzadas de múltiplas respostas 155
análise de sensibilidade
  na simulação 186
análise de série temporal
  prevendo casos 80
  previsão 80
análise de variância
  em Médias 25
  na Curva de Estimação 79
  na regressão linear 73
  no One-Way ANOVA 39
análise de what-if
  na simulação 186
Análise discriminante 97
  coeficientes de função 98
  critérios 99
  definindo intervalos 98
  Distância de Mahalanobis 99
  estatísticas 97, 98
  estatísticas descritivas 98
  example 97
  exportando informações de modelo 100
  gráficos 99
  Lamba de Wilks 99
  matriz de covariâncias 99
  matrizes 98
  métodos discriminantes 99
  métodos stepwise 97
  opções de exibição 99
  probabilidades anteriores 99
  recursos adicionais do comando 100
Análise discriminante *(continuação)*
  salvando variáveis de ordenação 100
  selecionando casos 98
  V de Rao 99
  valores omissos 99
  variáveis de agrupamento 97
  variáveis independentes 97
Análise do vizinho mais próximo 87
  opções 92
  partições 90
  saída 91
  salvando variáveis 91
  seleção de variável 90
  visualização do modelo 92
  vizinhos 89
Análise fatorial 103
  carregando gráficos 105
  convergência 104, 105
  descritiva 104
  escores dos fatores 106
  estatísticas 103, 104
  example 103
  formato de exibição de coeficiente 106
  métodos de extração 104
  métodos de rotação 105
  recursos adicionais do comando 106
  selecionando casos 104
  valores omissos 106
  visão geral 103
ANOVA
  em Médias 25
  em modelos lineares 66
  model 44
  no GLM Univariate 43
  no One-Way ANOVA 39
assimetria
  em Cubos OLAP 29
  em Descritivos 9
  em Frequências 5
  em Médias 25
  em Sumarizações de Relatórios em Colunas 162
  em Sumarizações de Relatórios em Linhas 160
  emSumarizar 22
  no Explore 12
associação linear por linear
  em Tabulações Cruzadas 16
autovalores
  na Análise fatorial 104
  na regressão linear 73

B

bagging
  em modelos lineares 61
Bonferroni
  no GLM 48
  no One-Way ANOVA 40

boosting
em modelos lineares 61
boxplots
comparando variáveis 12
comparar níveis de fator 12
na simulação 189
no Explore 12

C

C de Dunnett
no GLM 48
no One-Way ANOVA 40
carregando gráficos
na Análise fatorial 105
categoria de referência
no GLM 46
classificação
em Curva ROC 175
clusterização 112
escolhendo um procedimento 107
exibição geral 112
visualizando clusters 112
coeficiente alfa
na Análise de confiabilidade 165, 166
coeficiente de concordância de Kendall (W)
Testes Não Paramétricos de Amostras Relacionadas 134
coeficiente de contingência
em Tabulações Cruzadas 16
coeficiente de correlação de posto
em Correlações Bivariadas 55
coeficiente de correlação de Spearman
em Correlações Bivariadas 55
em Tabulações Cruzadas 16
coeficiente de correlação intraclasse (ICC)
na Análise de confiabilidade 166
coeficiente de correlação r
em Correlações Bivariadas 55
em Tabulações Cruzadas 16
coeficiente de dispersão (COD)
em Estatísticas de Razão 173
coeficiente de incerteza
em Tabulações Cruzadas 16
coeficiente de variação (COV)
em Estatísticas de Razão 173
coeficientes betas
na regressão linear 73
coeficientes de regressão
na regressão linear 73
combinações
em modelos lineares 63
comparações entre pares
testes não paramétricos 140
comparações múltiplas
no One-Way ANOVA 40
comparando grupos
em Cubos OLAP 31
comparando variáveis
em Cubos OLAP 31
confiabilidade de divisão de metade
na Análise de confiabilidade 165, 166
confiabilidade de Spearman-Brown
na Análise de confiabilidade 166
conjuntos de múltiplas respostas
Livro de códigos 1
Construtor de Simulação 180
contagem esperada
em Tabulações Cruzadas 18
contagem observada
em Tabulações Cruzadas 18
contrastes
no GLM 46
no One-Way ANOVA 39
contrastes de desvio
no GLM 46
contrastes de diferença
no GLM 46
contrastes de Helmert
no GLM 46
contrastes polinomiais
no GLM 46
no One-Way ANOVA 39
contrastes repetidos
no GLM 46
contrastes simples
no GLM 46
controle de página
em relatórios sumarização em colunas 163
em relatórios sumarização em linha 160
convergência
na Análise de Cluster por K-Médias 124
na Análise fatorial 104, 105
correção de Yates para continuidade
em Tabulações Cruzadas 16
Correlação de Pearson
em Correlações Bivariadas 55
em Tabulações Cruzadas 16
correlações
em Correlações Bivariadas 55
em Correlações Parciais 57
em Tabulações Cruzadas 16
na simulação 186
ordem zero 57
Correlações Bivariadas
coeficientes de correlação 55
estatísticas 55
nível de significância 55
opções 55
recursos adicionais do comando 56
valores omissos 55
correlações de ordem zero
em Correlações Parciais 57
Correlações parciais 57
correlações de ordem zero 57
estatísticas 57
na regressão linear 73
opções 57
recursos adicionais do comando 58
valores omissos 57
critério de informações de Akaike
em modelos lineares 63
critério de prevenção ao super ajuste
em modelos lineares 63
critérios de informações
em modelos lineares 63
Crosstabs 15
estatísticas 16
estratos 16
exibição da célula 18
Crosstabs *(continuação)*
formatos 19
gráficos de barras em cluster 16
suprimindo tabelas 15
variáveis de controle 16
Cubos OLAP 29
estatísticas 29
títulos 32
curtose
em Cubos OLAP 29
em Descritivos 9
em Frequências 5
em Médias 25
em Sumarizações de Relatórios em Colunas 162
em Sumarizações de Relatórios em Linhas 160
emSumarizar 22
no Explore 12
Curva de Estimação 79
análise de variância 79
incluindo constante 79
modelos 80
previsão 80
salvando intervalos de predição 80
salvando resíduos 80
salvando valores preditos 80
Curva ROC 175
estatísticas e gráficos 175

D

d
em Tabulações Cruzadas 16
d de Somers
em Tabulações Cruzadas 16
decomposição hierárquica 45
Definir conjuntos de múltiplas respostas 153
categorias 153
dicotomias 153
nomes de conjunto 153
rótulos do conjunto 153
dendrogramas
em Análise de Cluster Hierárquica 120
Descritivos 9
estatísticas 9
ordem de exibição 9
recursos adicionais do comando 10
salvando escores z 9
desvio médio absoluto (AAD)
em Estatísticas de Razão 173
desvio padrão
em Cubos OLAP 29
em Descritivos 9
em Estatísticas de Razão 173
em Frequências 5
em Médias 25
em Sumarizações de Relatórios em Colunas 162
em Sumarizações de Relatórios em Linhas 160
emSumarizar 22
no Explore 12
no GLM Univariate 47, 50, 52

DfBeta
na regressão linear 71
DfFit
na regressão linear 71
dicionário
Livro de códigos 1
diferença menos significativa
no GLM 48
no One-Way ANOVA 40
diferença significativa honesta de Tukey
no One-Way ANOVA 40
Diferença significativa honesta de Tukey
no GLM 48
diferenças entre grupos
em Cubos OLAP 31
diferenças entre variáveis
em Cubos OLAP 31
diferencial relacionado a preços (PRD)
em Estatísticas de Razão 173
distância de bloco
em Distâncias 59
distância de Chebychev
em Distâncias 59
distância de City Block
na Análise do vizinho mais próximo 89
Distância de Cook
na regressão linear 71
no GLM 51
Distância de Mahalanobis
na Análise Discriminante 99
na regressão linear 71
distância de Manhattan
na Análise do vizinho mais próximo 89
distância de Minkowski
em Distâncias 59
Distância euclidiana
em Distâncias 59
na Análise do vizinho mais próximo 89
distância euclidiana quadrática
em Distâncias 59
distância qui-quadrado
em Distâncias 59
Distâncias 59
calculando distâncias entre as variáveis 59
calcular distâncias entre casos 59
estatísticas 59
example 59
medidas de dissimilaridade 59
medidas de similaridade 60
medidas de transformação 59, 60
recursos adicionais do comando 60
valores de transformação 59, 60
distâncias de vizinho mais próximo
na Análise do vizinho mais próximo 94
diversas comparações post hoc 40
divisão
dividindo entre colunas do relatório 162

E

eliminação backward
na regressão linear 70
erro padrão
em Curva ROC 175
em Descritivos 9
em Frequências 5
no Explore 12
no GLM 47, 50, 51, 52
erro padrão da média
em Cubos OLAP 29
em Médias 25
emSumarizar 22
erro padrão de assimetria
em Cubos OLAP 29
em Médias 25
emSumarizar 22
erro padrão de curtose
em Cubos OLAP 29
em Médias 25
emSumarizar 22
escala
na Análise de confiabilidade 165
no Ajuste de escala multidimensional 169
Escala multidimensional 169
ajustando a escala de modelos 170
condicionalidade 170
criando matrizes de distância 170
critérios 171
definição de forma de dados 170
dimensões 170
estatísticas 169
exemplo 169
medidas de distância 170
níveis de medida 170
opções de exibição 171
recursos adicionais do comando 171
valores de transformação 170
escores dos fatores 106
escores dos fatores de Anderson-Rubin 106
escores dos fatores de Bartlett 106
escores z
em Descritivos 9
salvando como variáveis 9
estatística de Brown-Forsythe
no One-Way ANOVA 41
estatística de Cochran
em Tabulações Cruzadas 16
estatística de Durbin-Watson
na regressão linear 73
estatística de Mantel-Haenszel
em Tabulações Cruzadas 16
estatística de Welch
no One-Way ANOVA 41
estatística F
em modelos lineares 63
estatística R
em Médias 25
na regressão linear 73
estatísticas de proporções da coluna
em Tabulações Cruzadas 18
Estatísticas de razão 173
estatísticas 173
estatísticas descritivas
em Descritivos 9
estatísticas descritivas *(continuação)*
em Estatísticas de Razão 173
em Frequências 5
emSumarizar 22
na Análise de Cluster TwoStep 111
no Explore 12
no GLM Univariate 47, 50, 52
estimador de biponderação de Tukey
no Explore 12
estimador de onda de Andrews
no Explore 12
estimador M de Huber
no Explore 12
estimador-M redescendente de Hampel
no Explore 12
Estimadores M
no Explore 12
estimativas de efeito-tamanho
no GLM Univariate 47, 50, 52
estimativas de Hodges-Lehman
Testes Não Paramétricos de Amostras Relacionadas 134
estimativas de potência
no GLM Univariate 47, 50, 52
estimativas paramétrica
em Regressão Ordinal 76
no GLM Univariate 47, 50, 52
estratos
em Tabulações Cruzadas 16
estresse
no Ajuste de escala multidimensional 169
estresse de S
no Ajuste de escala multidimensional 169
estudo de caso de controle
Teste-T de amostras em pares 34
estudo de correspondência de pares
em Teste T de Amostras Emparelhadas 34
eta
em Médias 25
em Tabulações Cruzadas 16
eta quadrado
em Médias 25
no GLM Univariate 47, 50, 52
Explorar 11
estatísticas 12
gráficos 12
opções 13
recursos adicionais do comando 13
transformações de potência 12
valores omissos 13

F

F múltiplo de Ryan-Einot-Gabriel-Welsch
no One-Way ANOVA 40
fator de inflação de variância
na regressão linear 73
fatoração alfa 104
fatoração de eixo principal 104
fatoração de imagem 104
fi
em Tabulações Cruzadas 16
formatação
colunas em relatórios 160

forward stepwise
em modelos lineares 63
Frequências 5
estatísticas 5
formatos 7
gráficos 7
ordem de exibição 7
suprimindo tabelas 7
frequências acumulativas
em Regressão Ordinal 76
frequências de cluster
na Análise de Cluster TwoStep 111
Frequências de múltiplas respostas 154
valores omissos 154
frequências esperadas
em Regressão Ordinal 76
frequências observadas
em Regressão Ordinal 76
funções de densidade de probabilidade
na simulação 188
funções de distribuição cumulativas
na simulação 188

G

gama
em Tabulações Cruzadas 16
gama de Goodman e Kruskal
em Tabulações Cruzadas 16
GLM
gráficos de perfil 47
model 44
salvando matrizes 51
salvando variáveis 51
soma dos quadrados 44
testes post hoc 48
GLM Univariate 43, 48, 50, 52
contrastes 46
exibição 47, 50, 52
informações de diagnóstico 47, 50, 52
médias marginais estimadas 47, 50, 52
opções 47, 50, 52
gráfico de dispersão
na simulação 189
gráfico de espaço da variável
na Análise do vizinho mais próximo 92
gráfico icicle
em Análise de Cluster Hierárquica 120
gráficos
em Curva ROC 175
rótulos case 79
Gráficos de barra.
em Frequências 7
gráficos de dispersão
na regressão linear 71
gráficos de dispersão versus nível
no Explore 12
no GLM Univariate 47, 50, 52
gráficos de perfil
no GLM 47
gráficos de probabilidade normal
na regressão linear 71
no Explore 12
gráficos de ramos e folhas
no Explore 12
gráficos de resíduos
no GLM Univariate 47, 50, 52
gráficos de setores circulares
em Frequências 7
gráficos de tendência normal.
no Explore 12
gráficos de tornado
na simulação 189
gráficos parciais
na regressão linear 71
GT2 de Hochberg
no GLM 48
no One-Way ANOVA 40

H

H de Kruskal-Wallis
em Testes de duas amostras independentes 148
histogramas
em Frequências 7
na regressão linear 71
no Explore 12
histórico de iteração
em Regressão Ordinal 76

I

ICC. Consulte coeficiente de correlação intraclasse 166
importância de variável
na Análise do vizinho mais próximo 94
importância do preditor
modelos lineares 65
índice de concentração
em Estatísticas de Razão 173
informação de campo categórico
testes não paramétricos 140
informação de campo contínuo
testes não paramétricos 140
informações de diagnóstico de colinearidade
na regressão linear 73
informações de diagnóstico entre casos
na regressão linear 73
intervalo
em Cubos OLAP 29
em Descritivos 9
em Estatísticas de Razão 173
em Frequências 5
em Médias 25
emSumarizar 22
Intervalos de Clopper-Pearson
Testes Não paramétricos de uma amostra 128
intervalos de confiança
em Curva ROC 175
em Teste T de Amostras Emparelhadas 35
em Teste T de Amostras Independentes 34
em Teste T de Uma Amostra 36
na regressão linear 73
intervalos de confiança *(continuação)*
no Explore 12
no GLM 46, 47, 50, 52
no One-Way ANOVA 41
salvando na Regressão Linear 71
Intervalos de Jeffreys
Testes Não paramétricos de uma amostra 128
intervalos de predição
salvando em Curva de Estimação 80
salvando na Regressão Linear 71
intervalos de razão de verossimilhança
Testes Não paramétricos de uma amostra 128
iterações
na Análise de Cluster por K-Médias 124
na Análise fatorial 104, 105

K

k e seleção de variável
na Análise do vizinho mais próximo 95
kappa
em Tabulações Cruzadas 16
kappa de Cohen
em Tabulações Cruzadas 16
KR20
na Análise de confiabilidade 166
Kuder-Richardson 20 (KR20)
na Análise de confiabilidade 166

L

Lamba de Wilks
na Análise Discriminante 99
lambda
em Tabulações Cruzadas 16
lambda de Goodman e Kruskal
em Tabulações Cruzadas 16
limite inicial
na Análise de Cluster TwoStep 110
link
em Regressão Ordinal 76
listando casos 21
Livro de códigos 1
estatísticas 3
saída 1
LSD de Fisher
no GLM 48

M

mapa de quadrante
na Análise do vizinho mais próximo 94
matriz de correlações
em Regressão Ordinal 76
na Análise Discriminante 98
na Análise fatorial 103, 104
matriz de covariâncias
em Regressão Ordinal 76
na Análise Discriminante 98, 99
na regressão linear 73
no GLM 51

matriz de padrão
na Análise fatorial 103
matriz de transformação
na Análise fatorial 103
máxima verossimilhança
na Análise fatorial 104
máximo
comparando colunas de relatório 162
em Cubos OLAP 29
em Descritivos 9
em Estatísticas de Razão 173
em Frequências 5
em Médias 25
emSumarizar 22
no Explore 12
máximo de ramificações
na Análise de Cluster TwoStep 110
média
de diversas colunas de relatório 162
em Cubos OLAP 29
em Descritivos 9
em Estatísticas de Razão 173
em Frequências 5
em Médias 25
em Sumarizações de Relatórios em Colunas 162
em Sumarizações de Relatórios em Linhas 160
emSumarizar 22
no Explore 12
no One-Way ANOVA 41
subgrupo 25, 29
média aparada
no Explore 12
média geométrica
em Cubos OLAP 29
em Médias 25
emSumarizar 22
média harmônica
em Cubos OLAP 29
em Médias 25
emSumarizar 22
média ponderada
em Estatísticas de Razão 173
mediana
em Cubos OLAP 29
em Estatísticas de Razão 173
em Frequências 5
em Médias 25
emSumarizar 22
no Explore 12
mediana agrupada
em Cubos OLAP 29
em Médias 25
emSumarizar 22
Médias 25
estatísticas 25
opções 25
médias de grupo 25, 29
médias de subgrupo 25, 29
médias marginais estimadas
no GLM Univariate 47, 50, 52
médias observadas
no GLM Univariate 47, 50, 52
medida de diferença de tamanho
em Distâncias 59
medida de diferença padrão
em Distâncias 59
medida de dissimilaridade de Lance e Williams 59
em Distâncias 59
medida de distância fi-quadrado
em Distâncias 59
medidas de dispersão
em Descritivos 9
em Estatísticas de Razão 173
em Frequências 5
no Explore 12
medidas de distância
em Análise de Cluster Hierárquica 119
em Distâncias 59
na Análise do vizinho mais próximo 89
medidas de distribuição
em Descritivos 9
em Frequências 5
medidas de similaridade
em Análise de Cluster Hierárquica 119
em Distâncias 60
medidas de tendência central
em Estatísticas de Razão 173
em Frequências 5
no Explore 12
melhores subconjuntos
em modelos lineares 63
mínimo
comparando colunas de relatório 162
em Cubos OLAP 29
em Descritivos 9
em Estatísticas de Razão 173
em Frequências 5
em Médias 25
emSumarizar 22
no Explore 12
moda
em Frequências 5
modelagem espacial 197
modelagem geo-espacial 197
modelagem geoespacial 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208
modelo composto
na Curva de Estimação 80
modelo cúbico
na Curva de Estimação 80
modelo de crescimento
na Curva de Estimação 80
modelo de escala
em Regressão Ordinal 78
modelo de Guttman
na Análise de confiabilidade 165, 166
modelo de localização
em Regressão Ordinal 77
modelo de potência
na Curva de Estimação 80
modelo exponencial
na Curva de Estimação 80
modelo inverso
na Curva de Estimação 80
modelo linear
na Curva de Estimação 80
modelo logarítmico
na Curva de Estimação 80
modelo logístico
na Curva de Estimação 80
modelo paralelo
na Análise de confiabilidade 165, 166
modelo paralelo-série
na Análise de confiabilidade 165, 166
modelo quadrático
na Curva de Estimação 80
modelo S
na Curva de Estimação 80
modelos customizados
no GLM 44
modelos fatoriais completos
no GLM 44
modelos lineares 61
coeficientes 66
combinações 63
critério de informações 64
estatística R-quadrado 64
importância do preditor 65
médias estimadas 67
nível de confiança 62
objetivos 61
opções de modelo 64
predito por observado 65
preparação de dado automático 62, 64
regras de combinação 63
replicando resultados 64
resíduos 65
seleção de modelo 63
sumarização de construção de modelo 67
sumarização do modelo 64
Tabela de ANOVA 66
valores discrepantes 66
Múltiplas respostas
recursos adicionais do comando 157
multiplicação
multiplicando entre colunas do relatório 162
Múltiplos F de Ryan-Einot-Gabriel-Welsch
no GLM 48

N

Newman-Keuls
no GLM 48
numeração de página
em relatórios sumarização em colunas 163
em relatórios sumarização em linha 160
número de casos
em Cubos OLAP 29
em Médias 25
emSumarizar 22

O

One-Way ANOVA 39
comparações múltiplas 40
contrastes 39
contrastes polinomiais 39

One-Way ANOVA *(continuação)*
estatísticas 41
opções 41
recursos adicionais do comando 42
testes post hoc 40
valores omissos 41
variáveis de fator 39

P

padronização
na Análise de Cluster TwoStep 110
peers
na Análise do vizinho mais próximo 94
percentis
em Frequências 5
na simulação 189
no Explore 12
PLUM
em Regressão Ordinal 75
porcentagens
em Tabulações Cruzadas 18
porcentagens de coluna
em Tabulações Cruzadas 18
porcentagens de linha
em Tabulações Cruzadas 18
porcentagens totais
em Tabulações Cruzadas 18
preparação de dado automático
em modelos lineares 64
previsão
na Curva de Estimação 80
primeiro
em Cubos OLAP 29
em Médias 25
emSumarizar 22
profundidade da árvore
na Análise de Cluster TwoStep 110
Proximidades
em Análise de Cluster Hierárquica 119

Q

Q de Cochran
em Testes para várias amostras relacionadas 150
quadrados mínimos generalizados
na Análise fatorial 104
quadrados mínimos não ponderados
na Análise fatorial 104
quadrados mínimos ponderados
na regressão linear 69
qualidade do ajuste
em Regressão Ordinal 76
quartis
em Frequências 5
qui-quadrado 141
associação linear por linear 16
correção de Yates para continuidade 16
em Tabulações Cruzadas 16
estatísticas 142
intervalo esperado 142
opções 142
qui-quadrado *(continuação)*
para a independência 16
Pearson 16
razão de verossimilhança 16
teste de uma amostra 141
teste exato de Fisher 16
valores esperados 142
valores omissos 142
Qui-quadrado de Pearson
em Regressão Ordinal 76
em Tabulações Cruzadas 16
qui-quadrado de razão de verossimilhança
em Regressão Ordinal 76
em Tabulações Cruzadas 16

R

R 2
em Médias 25
mudança de R 2 73
na regressão linear 73
R 2 ajustado
na regressão linear 73
R-E-G-W F
no GLM 48
no One-Way ANOVA 40
R-E-G-W Q
no GLM 48
no One-Way ANOVA 40
R múltiplo
na regressão linear 73
R-quadrado
em modelos lineares 64
R quadrado ajustado
em modelos lineares 63
R2 de Cox e Snell
em Regressão Ordinal 76
R2 de McFadden
em Regressão Ordinal 76
R2 de Nagelkerke
em Regressão Ordinal 76
razão de covariância
na regressão linear 71
regras de combinação
em modelos lineares 63
regressão
gráficos 71
Regressão linear 69
regressão múltipla 69
Regressão linear 69
Blocos. 69
estatísticas 73
exportando informações de modelo 71
gráficos 71
métodos de seleção de variáveis 70, 73
pesos 69
recursos adicionais do comando 74
resíduos 71
salvando novas variáveis 71
valores omissos 73
variável de seleção 70
regressão múltipla
na regressão linear 69
Regressão ordinal 75
Regressão ordinal *(continuação)*
estatísticas 75
link 76
modelo de escala 78
modelo de localização 77
opções 76
recursos adicionais do comando 78
Regressão por quadrados mínimos parciais 83
exportar variáveis 85
model 84
relatórios
comparando colunas 162
dividindo valores de coluna 162
multiplicando valores de coluna 162
relatórios sumarização em colunas 161
relatórios sumarização em linha 159
totais compostos 162
total de colunas 162
relatórios sumarização em colunas 161
resíduos
em Tabulações Cruzadas 18
salvando em Curva de Estimação 80
salvando na Regressão Linear 71
resíduos de Pearson
em Regressão Ordinal 76
resíduos Estudentizados
na regressão linear 71
resíduos excluídos
na regressão linear 71
no GLM 51
resíduos não padronizados
no GLM 51
resíduos padronizados
na regressão linear 71
no GLM 51
Resumir 21
estatísticas 22
opções 21
risco
em Tabulações Cruzadas 16
risco relativo
em Tabulações Cruzadas 16
rô
em Correlações Bivariadas 55
em Tabulações Cruzadas 16
rotação equamax
na Análise fatorial 105
rotação oblimin direta
na Análise fatorial 105
rotação quartimax
na Análise fatorial 105
rotação varimax
na Análise fatorial 105

S

seleção de variável
na Análise do vizinho mais próximo 95
seleção forward
na Análise do vizinho mais próximo 90
na regressão linear 70

seleção k
na Análise do vizinho mais próximo 95
seleção stepwise
na regressão linear 70
Sequências de Wald-Wolfowitz
em Testes de duas amostras independentes 146
simulação 177
ajuste de distribuição 183
amostragem de cauda 187
análise de sensibilidade 186
análise de what-if 186
box plots 189
Construtor de Simulação 180
correlações entre entradas 186
criando novas entradas 182
criando um plano de simulação 177, 178, 179
critérios de parada 187
customizando o ajuste de distribuição 185
editor de equação 181
especificação de modelo 180
executando um plano de simulação 179, 191
formatos de exibição para respostas e entradas 189
função de densidade de probabilidade 188
função de distribuição cumulativa 188
gráficos de dispersão 189
gráficos de tornado 189
gráficos interativos 193
modelos suportados 180
opções de gráfico 194
percentis de distribuição de resposta 189
reajustando as distribuições para novos dados 191
resultados do ajuste de distribuição 185
saída 188, 189
salvar dados simulados 190
salvar plano de simulação 190
simulação de Monte Carlo 177
sinalizar teste
em Testes de Duas Amostras Relacionadas 147
soma
em Cubos OLAP 29
em Descritivos 9
em Frequências 5
em Médias 25
emSumarizar 22
soma dos quadrados 45
no GLM 44
Student-Newman-Keuls
no GLM 48
no One-Way ANOVA 40
subconjuntos homogêneos
testes não paramétricos 140
subtotais
em relatórios sumarização em colunas 163
sumarização de erro
na Análise do vizinho mais próximo 95
sumarização de hipótese
testes não paramétricos 136
sumarização do intervalo de confiança
testes não paramétricos 136, 137
Sumarizações de Relatórios em Colunas 161
controle de página 163
formato de coluna 160
layout da página 161
numeração de página 163
recursos adicionais do comando 164
subtotais 163
total de colunas 162
total geral 163
valores omissos 163
Sumarizações de Relatórios em Linhas 159
colunas de dados 159
colunas de quebra 159
controle de página 160
espaçamento de quebra 160
formato de coluna 160
layout da página 161
numeração de página 160
recursos adicionais do comando 164
Rodapés. 161
sequências de ordenação 159
títulos 161
valores omissos 160
variáveis em títulos 161

T

T2 de Tamhane
no GLM 48
no One-Way ANOVA 40
T3 de Dunnett
no GLM 48
no One-Way ANOVA 40
tabela de classificação
na Análise do vizinho mais próximo 95
tabelas de contingência 15
tabelas de frequência
em Frequências 5
tabelas de frequências
no Explore 12
tabulação cruzada
em Tabulações Cruzadas 15
múltiplas respostas 155
Tabulações cruzadas de múltiplas respostas 155
correspondendo variáveis em conjuntos de respostas 156
definindo intervalos de valor 156
porcentagens com base em casos 156
porcentagens com base em respostas 156
porcentagens de célula 156
valores omissos 156
tau-b
em Tabulações Cruzadas 16
Tau-b de Kendall
em Correlações Bivariadas 55
Tau-b de Kendall *(continuação)*
em Tabulações Cruzadas 16
tau-c
em Tabulações Cruzadas 16
Tau-c de Kendall 16
em Tabulações Cruzadas 16
tau de Goodman e Kruskal
em Tabulações Cruzadas 16
tau de Kruskal
em Tabulações Cruzadas 16
termos de compilação 45, 77, 78
termos de interação 45, 77, 78
teste b de Tukey
no One-Way ANOVA 40
Teste b de Tukey
no GLM 48
teste binomial
Testes Não paramétricos de uma amostra 128
Testes Não Paramétricos de Uma Amostra 128
Teste binomial
recursos adicionais do comando 143
Teste de aditividade de Tukey
na Análise de confiabilidade 165, 166
teste de amostras independentes
testes não paramétricos 138
teste de amplitude múltipla de Duncan
no GLM 48
no One-Way ANOVA 40
Teste de Binômio 142
dicotomias 142
estatísticas 143
opções 143
valores omissos 143
teste de comparação entre pares de Gabriel
no GLM 48
no One-Way ANOVA 40
teste de comparações pairwise de Games-Howell
no GLM 48
no One-Way ANOVA 40
teste de esfericidade de Bartlett
na Análise fatorial 104
teste de execuções
Testes Não Paramétricos de Uma Amostra 128, 129
Teste de Execuções
estatísticas 144
opções 144
pontos de corte 144
recursos adicionais do comando 144
valores omissos 144
teste de Friedman
Testes Não Paramétricos de Amostras Relacionadas 134
Teste de Friedman
em Testes para várias amostras relacionadas 150
teste de homogeneidade marginal
em Testes de Duas Amostras Relacionadas 147
Testes Não Paramétricos de Amostras Relacionadas 134

teste de Kolmogorov-Smirnov
Testes Não paramétricos de uma amostra 128, 129
Teste de Kolmogorov-Smirnov de uma amostra 145
distribuição de teste 145
estatísticas 145
opções 145
recursos adicionais do comando 145
valores omissos 145
teste de Levene
no Explore 12
no GLM Univariate 47, 50, 52
no One-Way ANOVA 41
teste de Lilliefors
no Explore 12
teste de linhas paralelas
em Regressão Ordinal 76
teste de McNemar
em Tabulações Cruzadas 16
em Testes de Duas Amostras Relacionadas 147
Testes Não Paramétricos de Amostras Relacionadas 134, 135
teste de mediana
em Testes de duas amostras independentes 148
Teste de reação extrema de Moses
em Testes de duas amostras independentes 146
teste de Scheffé
no One-Way ANOVA 40
Teste de Scheffé
no GLM 48
teste de Shapiro-Wilk
no Explore 12
Teste de Sidak
no GLM 48
teste de sinal
Testes Não Paramétricos de Amostras Relacionadas 134
teste dos postos sinalizados de Wilcoxon
em Testes de Duas Amostras Relacionadas 147
Testes Não Paramétricos de Amostras Relacionadas 134
Testes Não Paramétricos de Uma Amostra 128
teste exato de Fisher
em Tabulações Cruzadas 16
teste M de Box
na Análise Discriminante 98
teste Q de Cochran
Testes Não Paramétricos de Amostras Relacionadas 134, 135
teste qui-quadrado
Testes Não paramétricos de uma amostra 129
Testes Não Paramétricos de Uma Amostra 128
teste t
em Teste T de Amostras Emparelhadas 34
em Teste T de Amostras Independentes 33
em Teste T de Uma Amostra 35
no GLM Univariate 47, 50, 52
Teste-T de amostras em pares 34
opções 35
selecionando variáveis pairwise 34
valores omissos 35
Teste-T de amostras independentes 33
definindo grupos 34
intervalos de confiança 34
opções 34
valores omissos 34
variáveis de agrupamento 34
variáveis de sequência de caracteres 34
teste t de duas amostras
em Teste T de Amostras Independentes 33
teste t de Dunnett
no GLM 48
no One-Way ANOVA 40
teste t de Sidak
no One-Way ANOVA 40
teste t de Student 33
Teste-T de uma amostra 35
intervalos de confiança 36
opções 36
recursos adicionais do comando 35, 36
valores omissos 36
teste t de Waller-Duncan
no GLM 48
no One-Way ANOVA 40
teste t dependente
em Teste T de Amostras Emparelhadas 34
Testes de duas amostras independentes 146
definindo grupos 147
estatísticas 147
opções 147
recursos adicionais do comando 147
tipos de teste 146
valores omissos 147
variáveis de agrupamento 147
Testes de duas amostras relacionadas
estatísticas 148
opções 148
recursos adicionais do comando 148
tipos de teste 148
valores omissos 148
Testes de Duas Amostras Relacionadas 147
testes de homogeneidade de variância
no GLM Univariate 47, 50, 52
no One-Way ANOVA 41
testes de independência
qui-quadrado 16
testes de linearidade
em Médias 25
testes de normalidade
no Explore 12
testes não paramétricos
qui-quadrado 141
Teste de Execuções 144
Teste de Kolmogorov-Smirnov de uma amostra 145
Testes de duas amostras independentes 146
testes não paramétricos *(continuação)*
Testes de Duas Amostras Relacionadas 147
Testes para diversas amostras independentes 148
Testes para várias amostras relacionadas 150
visualização do modelo 136
Testes Não Paramétricos de Amostras Independentes 130
guia Campos 131
Testes Não Paramétricos de Amostras Relacionadas 133
campos 133
teste de McNemar 135
teste Q de Cochran 135
Testes Não paramétricos de uma amostra 127
campos 127
teste binomial 128
teste de execuções 129
teste de Kolmogorov-Smirnov 129
teste qui-quadrado 129
Testes para diversas amostras independentes 148
definindo intervalo 149
estatísticas 149
opções 149
recursos adicionais do comando 150
tipos de teste 149
valores omissos 149
variáveis de agrupamento 149
Testes para várias amostras relacionadas 150
estatísticas 150
recursos adicionais do comando 151
tipos de teste 150
títulos
em Cubos OLAP 32
tolerância
na regressão linear 73
totais gerais
em relatórios sumarização em colunas 163
total de coluna
em relatórios 162
tratamento de ruído
na Análise de Cluster TwoStep 110

U

U de Mann-Whitney
em Testes de duas amostras independentes 146
último
em Cubos OLAP 29
em Médias 25
emSumarizar 22

V

V
em Tabulações Cruzadas 16
V de Cramér
em Tabulações Cruzadas 16

V de Rao
na Análise Discriminante 99
valores de ponto de alavanca
na regressão linear 71
no GLM 51
valores discrepantes
na Análise de Cluster TwoStep 110
na regressão linear 71
no Explore 12
valores extremos
no Explore 12
valores omissos
em Correlações Bivariadas 55
em Correlações Parciais 57
em Crosstabs de Múltiplas Respostas 156
em Curva ROC 175
em Frequências de Múltiplas Respostas 154
em relatórios sumarização em colunas 163
em Sumarizações de Relatórios em Linhas 160
em Teste Binomial 143
em Teste de Execuções 144
em Teste de Kolmogorov-Smirnov de Uma Amostra 145
em Teste Qui-Quadrado 142
em Teste T de Amostras Emparelhadas 35
em Teste T de Amostras Independentes 34
em Teste T de Uma Amostra 36
em Testes de duas amostras independentes 147
em Testes de Duas Amostras Relacionadas 148
em Testes para Várias Amostras Independentes 149
na Análise do vizinho mais próximo 92
na Análise fatorial 106
na regressão linear 73
no Explore 13
no One-Way ANOVA 41
valores padronizados
em Descritivos 9
valores preditos
salvando em Curva de Estimação 80
salvando na Regressão Linear 71
valores preditos ponderados
no GLM 51
variação
em Cubos OLAP 29
em Descritivos 9
em Frequências 5
em Médias 25
em Sumarizações de Relatórios em Colunas 162
em Sumarizações de Relatórios em Linhas 160
emSumarizar 22
no Explore 12
variáveis de controle
em Tabulações Cruzadas 16
variável de seleção
na regressão linear 70
visualização
modelos de clusterização 112
visualização do modelo
na Análise do vizinho mais próximo 92
testes não paramétricos 136
visualizador de cluster
comparação de clusters 115
distribuição de células 115
exibição de conteúdo da célula 114
filtrando registros 116
importância do preditor 115
inverter clusters e variáveis 114
ordem de exibição da variável 114
ordem de exibição do cluster 114
ordenar clusters 114
ordenar conteúdo da célula 114
ordenar variáveis 114
sobre modelos de cluster 112
sumarização do modelo 113
tamanho de clusters 115
transpor clusters e variáveis 114
usando 115
visão geral 112
visualização básica 114
visualização de centros do cluster 113
visualização de clusters 113
visualização de comparação do cluster 115
visualização de distribuição de célula 115
visualização de importância do preditor de cluster 115
visualização de sumarização 113
visualização de tamanhos de cluster 115

W

W de Kendall
em Testes para várias amostras relacionadas 150

Z

Z Kolmogorov-Smirnov
em Teste de Kolmogorov-Smirnov de Uma Amostra 145
em Testes de duas amostras independentes 146

Impresso no Brasil