



Bioestatística quantitativa aplicada

Edison Capp
Otto Henrique Nienov
Organizadores

Caroline Darski
Charles Francisco Ferreira
Cristiana Palma Kuhl
Fernanda Dapper Machado
Fernanda Vargas Ferreira
Hellen Meiry Grosskopf Werka
Johanna Ovalle Diaz
Marina Petter Rodrigues
Michele Strelow Moreira
Nadine de Souza Ziegler
Paula Barros Terraciano
Pedro Henrique Comerlato
Sinara Santos

Universidade Federal do Rio Grande do Sul
Faculdade de Medicina
Programa de Pós-Graduação em Ciências da Saúde:
Ginecologia e Obstetrícia

Bioestatística Quantitativa Aplicada

Porto Alegre 2020
UFRGS

U58b Universidade Federal do Rio Grande do Sul. Faculdade de Medicina. Programa de Pós-Graduação em Ciências da Saúde: Ginecologia e Obstetrícia Bioestatística quantitativa aplicada/ Universidade Federal do Rio Grande do Sul; organizadores: Edison Capp e Otto Henrique Nienov – Porto Alegre: UFRGS, 2020.

260p.

ISBN: 978-65-86232-43-1

E-Book: 978-65-86232-44-8

1. Epidemiologia e Bioestatística 2. Estatística 3. SPSS I. Capp, Edison, org. II. Nienov, Otto Henrique, org. III Título.

NLM: WA950

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
(Bibliotecária Shirlei Galarça Salort – CRB10/1929)

Endereço:

PPG em Ciências da Saúde: Ginecologia e Obstetrícia

FAMED – UFRGS

Rua Ramiro Barcellos, 2400/2º andar

CEP 900035-003 – Porto Alegre – RS

Telefone: +55 51 3308 5607

E-mail: ppggo@ufrgs.br

Editoração e diagramação: Edison Capp

Capa: Edison Capp, imagens: www.freepik.com/starline

Edison Capp
Otto Henrique Nienov
Organizadores

Caroline Darski
Charles Francisco Ferreira
Cristiana Palma Kuhl
Fernanda Dapper Machado
Fernanda Vargas Ferreira
Hellen Meiry Grosskopf Werka
Johanna Ovalle Diaz
Marina Petter Rodrigues
Michele Strelow Moreira
Nadine de Souza Ziegler
Paula Barros Terraciano
Pedro Henrique Comerlato
Sinara Santos

4 Transformando e manipulando dados no SPSS

*Marina Petter Rodrigues
Michele Strelow Moreira
Edison Capp
Otto Henrique Nienov*

Por que transformar e manipular dados?

O banco de dados é uma estrutura que organiza e armazena informações relacionadas a unidade experimental (pessoa, animal ou objeto), obtidas através de um instrumento de coleta (como fichas de coleta, planilhas, prontuários, etc). Um banco de dados precisa ser planejado. Cada linha corresponde a uma unidade experimental (ou caso) e contém os dados sobre aquela unidade. Cada coluna corresponde a uma variável e contém um único tipo de informação que aparece para cada caso do estudo. Uma boa forma de elaborar o banco de dados é avaliar se as variáveis contidas nele são suficientes para responder os objetivos propostos pelo projeto de pesquisa.

Precisamos considerar todas as formas de avaliar os dados que serão coletados. No entanto, algumas vezes, os dados com os quais começamos nem sempre são organizados da melhor forma para sua análise. Por exemplo, você pode querer criar uma variável categórica a partir de uma variável quantitativa, ou criar uma variável de tempo a partir da diferença entre duas datas, ou combinar variáveis categóricas em uma única variável, ou criar uma variável através de um cálculo com outras variáveis existentes, ou, ainda, criar uma variável a partir da contagem de outras.

Além disso, há casos em que os arquivos de dados não estão organizados da forma ideal para sua análise. Por exemplo, você pode querer classificar os casos com base no valor de uma ou mais variáveis, ou restringir sua análise a um subconjunto de casos, ou executar análises simultâneas em diferentes subconjuntos, ou, ainda, incluir novos casos ou novas variáveis no banco de dados.

Nesse sentido, o SPSS fornece ferramentas de manipulação de dados integradas para gravação e transformação de variáveis.

Para demonstrar essas ferramentas, neste capítulo será utilizado o “Banco de dados 3.sav” (disponível em <https://bit.ly/bancosdedados>), que contém dados demográficos, antropométricos, de pressão arterial e perfil lipídico de uma amostra de sujeitos obesos. Além disso, os sujeitos foram classificados de acordo com a frequência e duração das atividades cotidianas (caminhadas, atividades físicas moderadas e vigorosas) realizadas nos últimos 7 dias em cinco categorias: sedentários, insuficientemente ativos B e A, ativos ou muito ativos, de acordo com o Questionário Internacional de Atividade Física (IPAQ) versão curta. Lembre-se: é importante conhecer o banco de dados. Tome algum tempo para examiná-lo e conhecer as variáveis.

Transformar variável quantitativa em categórica

Uma abordagem muito utilizada é a de categorizar as variáveis quantitativas, ou seja, transformar uma variável quantitativa em uma variável qualitativa ordinal. A categorização significa escolher dois ou mais valores contínuos e agrupá-los na mesma categoria. A transformação de variáveis quantitativas em categóricas pode ser utilizada para facilitar a caracterização da amostra.

Para exemplificar, iremos categorizar a variável “circunferência da cintura” (“CC”) utilizando os quartis. Lembre-se que os quartis são pontos de corte na escala da variável de tal forma que, cada grupo formado a partir destes pontos de corte terá 1/4 dos casos, ou seja, 25% do tamanho total da amostra. Podemos categorizar variáveis quantitativas de diferentes formas.

Uma das formas de categorização é utilizando a opção “Recodificar em variáveis diferentes”, do menu “Transformar”. Para isso, primeiro precisamos calcular os quartis para definirmos os pontos de corte. Como visto no capítulo 3, obtemos os quartis no menu “Analisar”, clicando em “Estatísticas descritivas” e, em

seguida, em "Frequências". Na janela "Frequências", localize na tela da esquerda a variável de interesse (por exemplo, "circunferência da cintura") e clique na seta para mover para a lista de "Variável(is)". Clique em "Estatísticas" e assinale a opção "Quartis". Clique em "Continuar" e, após, em "OK" ou em "Colar" para gerar o arquivo de sintaxe. Lembre-se que, se você colar a sintaxe, é preciso selecioná-la para rodar o comando clicando em "Executar seleção". Os resultados da análise aparecerão no arquivo de saída e, a partir dos valores dos quartis obtidos, iremos categorizar a variável "circunferência da cintura" conforme mostrado no quadro 1. O número de categorias de agrupamento é um número maior do que o número de pontos de corte. Portanto, neste exemplo, a nova variável categorizada terá quatro categorias, sendo que cada uma irá conter 25% do tamanho total da amostra.

Quadro 1. Pontos de corte para categorização da variável "circunferência da cintura".

CATEGORIA	INTERVALO DE VALORES
1	Até 104,00 cm
2	104,01 cm a 111,00 cm
3	111,01 cm a 120,00 cm
4	Mais de 120,01 cm

Definidos os pontos de corte, vamos categorizar a variável. Clique em "Transformar" e, em seguida, em "Recodificar em variáveis diferentes". Na janela "Recodificar em variáveis diferentes", localize na tela da esquerda a variável "Circunferência da cintura" e clique na seta para mover para a tela "Variável numérica -> Variável de saída". Atribua um novo nome para a variável de saída, como por exemplo "CCREC", em "Nome". Informe também uma descrição para a variável que será criada, como por exemplo "Circunferência da cintura recodificada", em "Rótulo". Clique em "Alterar". Criada a nova variável, precisamos informar as quatro categorias definidas pelos valores dos quartis. Para isso, clique em "Valores antigo e novo". Clique em "Faixa, MAIS BAIXO até o valor" e digite o valor obtido para o primeiro

quartil, no caso "104,00", que corresponderá ao menor valor de corte. Em "Novo valor", digite "1", pois, conforme apresentado no quadro 1, é a primeira categoria. Clique em "Adicionar". Na tela "Antigo -> Novo" estará indicado o valor correspondente a categoria 1. Em seguida, vamos informar nossa segunda categoria. Como corresponde a uma faixa de valores, clique em "Faixa", e indique o intervalo do valor, neste caso, "104,01" até "111,00". Em "Novo valor", digite "2", pois corresponde a nossa segunda categoria. Clique em Adicionar. Na tela "Antigo -> Novo" estará indicado o valor correspondente a categoria 2. Repita a ação, indicando na "Faixa" o intervalo da terceira categoria, neste caso, "111,01" até "120,00". E, em "Novo valor", digite "3". Clique em "Adicionar". Por fim, precisamos indicar o último ponto de corte. Selecione a opção "Faixa, valor até o MAIS ALTO" e digite o valor "120,01". Em "Novo valor", digite "4". Clique em "Adicionar". Pronto! As quatro categorias com os respectivos pontos de corte foram informadas. Clique em "Continuar" e, em seguida, em "Ok" ou em "Colar" para gerar o arquivo de sintaxe. Lembre-se que, se você colar a sintaxe, é preciso selecioná-la para rodar o comando clicando em "Executar seleção". A nova variável "CCREC", correspondente à variável "CC" categorizada, automaticamente será incluída no final do banco de dados. É importante informar os valores dos rótulos da variável criada em "Valores" na planilha "Visualização da variável".

Outra forma de categorizar uma variável quantitativa é através da opção "Armazenamento visual", no menu "Transformar". Na janela de "Armazenamento visual", localize na tela da esquerda a variável de interesse (por exemplo, "circunferência da cintura") e mova para a lista "Variáveis para armazenar". Clique em "Continuar". Na janela seguinte, em "Variável armazenada", defina um novo nome para a variável, como por exemplo "CCQUARTIS", pois também vamos categorizá-la baseado nos valores dos quartis. Perceba que há um histograma que exhibe a distribuição da variável selecionada. Clique em "Fazer pontos de corte". Escolha a opção "Percentis iguais baseados em casos verificados". Preencha uma das seguintes opções: ou "Número de pontos de corte:" n° de grupos (categorias) - 1, no caso, $4 - 1 = "3"$; ou "Largura (%):" % de casos em cada grupo, no caso como

são quatro grupos de mesmo tamanho, $100/4 = "25"$. Clique em "Aplicar". Os valores exibidos na grade representam os pontos de corte definidos, que são os valores terminais superiores de cada categoria. As linhas verticais no histograma também indicam os locais dos pontos de corte. Por padrão, esses valores de ponto de corte são incluídos nas categorias correspondentes. Por exemplo, a primeira categoria irá incluir todos os valores menores ou iguais a 104,00 cm. É possível inserir os valores dos rótulos automaticamente clicando em "Fazer rótulos". Isso gera automaticamente rótulos de valores descritivos para cada categoria. Você também pode inserir manualmente ou alterar pontos de corte e rótulos diretamente na grade, alterar locais de pontos de corte arrastando e soltando as linhas dos pontos de corte no histograma e excluir pontos de corte arrastando as linhas de pontos de corte para fora do histograma. Clique em "Ok" ou em "Colar" para gerar o arquivo de sintaxe. Uma mensagem de alerta indicará que "As especificações de armazenamento criarão 1 variável". Clique em "Ok". A nova variável "CCQUARTIS" é exibida na janela de edição de dados. Como a variável é incluída no final do arquivo, ela será exibida na coluna à extrema direita na planilha "Visualização de Dados" e na última linha na planilha "Visualização de Variável".

Calcular a diferença entre duas datas

Uma das tarefas mais comuns envolvendo datas é calcular o período de tempo entre duas datas. Para exemplificar, iremos criar uma variável "Idade" a partir das variáveis de "Data de nascimento" ("DN") e de "Data da coleta de dados" ("DATACOL"), utilizando a função "Datediff (datetime2,datetime1,"unit")". Para isso, no menu "Transformar", clique em "Calcular variável...". Na janela "Calcular variável", em "Variável de destino", digite o nome da nova variável, como por exemplo "IDADE". Clique em "Tipo e rótulo" e defina uma expressão para a variável, como "Idade (em anos)", ou utilize a própria expressão como rótulo, clicando em "Usar expressão como rótulo". Na tela "Grupo de função", selecione "Aritmética de data" e, na tela abaixo ("Funções e variáveis especiais"), selecione a opção "Datediff". Ao selecionar a opção, clique na seta para inseri-la na tela de

"Expressão numérica". A função será inserida na expressão. A função não é concluída até você inserir os argumentos, representados por pontos de interrogação na função colada. O número de pontos de interrogação indica o número mínimo de argumentos requeridos para concluir a função. Leia atentamente a descrição da função "*Datediff (datetime2,datetime1,\"unit\")*": será calculada a diferença entre a data da coleta ("*datetime2*") e a data de nascimento ("*datetime1*") para obtermos a idade em anos ("*years*"). Desta forma, destaque o primeiro ponto de interrogação na função colada, localize na tela da esquerda a variável "Data da coleta de dados" e clique na seta para inseri-la na "Expressão numérica" (a variável selecionada deve ficar no lugar de "*datetime2*"). Em seguida, destaque o segundo ponto de interrogação na função colada, localize na mesma tela a variável "Data de nascimento" e clique na seta para inseri-la (a variável selecionada deve ficar no lugar de "*datetime1*"). Em "*units*" (que corresponde ao terceiro ponto de interrogação), digite, entre aspas, a unidade de tempo em que a variável será expressa, no caso, "*years*" (utilize "*months*" para meses, "*days*" para dias, etc.). Clique em "Ok" ou em "Colar" para gerar o arquivo de sintaxe. A variável será incluída no final do banco de dados.

Combinando variáveis categóricas

Em algumas situações pode ser interessante combinar duas variáveis categóricas para obtermos uma variável categórica combinada. Para exemplificar, iremos combinar as variáveis "Sexo" ("*SEXO*") e "Hipertensão" ("*HAS*") para criarmos a variável "*SEXOHAS*". Sabendo que a variável "Sexo" é categorizada da seguinte forma: 1 = Sexo masculino e 2 = Sexo feminino e, a variável "Hipertensão" é categorizada em: 0 = Não hipertenso e 1 = Hipertenso, a variável "*SEXOHAS*" poderá apresentar as categorias apresentadas no Quadro 2.

Quadro 2. Combinação das categorias das variáveis “SEXO” e “HAS” para criação das categorias da variável “SEXOHAS”.

CATEGORIA	VALOR DOS RÓTULOS
1	Sexo masculino hipertenso
2	Sexo masculino não hipertenso
3	Sexo feminino hipertenso
4	Sexo feminino não hipertenso

No menu “Transformar”, clique em “Calcular variável...”. Lembre-se que é possível redefinir os comandos de teste na opção “Redefinir”. Em seguida, na janela “Calcular variável”, informe o nome da nova variável, como por exemplo “SEXOHAS”, em “Variável de destino”. Clique em “Tipo e rótulo” e defina uma expressão para a variável, como “Combinação de sexo e hipertensão”, ou utilize a própria expressão como rótulo, clicando em “Usar expressão como rótulo”. Na tela “Expressão numérica”, digite “1”. Em seguida, vamos informar a combinação de valores de rótulos que define a primeira categoria da variável, conforme apresentado no quadro 2. Clique em “Se...(condição de seleção de caso opcional)”. Selecione a opção “Incluir se o caso atender à condição”. Em seguida, localize na tela da esquerda a variável “SEXO”, selecione e insira a variável dentro da tela “Incluir se o caso atender à condição”. Digite ao lado da expressão SEXO utilizando o teclado virtual localizado no centro da janela, “= 1” (correspondendo ao “Sexo masculino”). Agora, precisamos inserir o restante da combinação que define a primeira categoria. Insira “&” e localize na tela da esquerda a variável “HAS”. Selecione e insira a variável dentro da tela “Incluir se o caso atender à condição”. Na mesma tela, ao lado da expressão HAS, digite utilizando o teclado virtual localizado no centro da janela, “= 1” (correspondendo ao “Hipertenso”). Após esse procedimento a expressão na janela deve ser a seguinte: “SEXO = 1 & HAS = 1”, ou seja, a categoria 1 = Sexo masculino hipertenso. Clique em “Continuar”. Em seguida, clique em “Ok ou em “Colar” para gerar o arquivo de sintaxe. A variável será incluída no final do banco de dados.

No entanto, temos somente a primeira categoria da nova variável informada. Ainda precisamos incluir as outras três categorias da variável "SEXOHAS". Para isso, procedese de maneira análoga, alterando o código na tela "Expressão numérica" para "2", "3" e "4" e a expressão da tela "Incluir se o caso atender à condição" para "SEXO = 1 e HAS = 0", "SEXO = 2 e HAS = 1" e "SEXO = 2 e HAS = 0", respectivamente.

Na janela de edição de dados existe um atalho onde é possível acessar de forma rápida os comandos acessados recentemente, em "Chamar novamente as caixas de diálogo usadas recentemente". Caso contrário, clique novamente em "Calcular variável...", no menu "Transformar". Na tela "Expressão numérica", digite "2". Clique em "Se..." para alterar a expressão. Na tela "Incluir se o caso atender à condição", informe utilizando o teclado virtual localizado no centro da janela, a expressão "SEXO = 1 & HAS = 0", ou seja, a categoria 2 = Sexo masculino não hipertenso. Clique em "Continuar". Em seguida, clique em "Ok" ou em "Colar" para gerar o arquivo de sintaxe. Uma mensagem perguntará se você deseja "Alterar variável existente?". Clique em "OK". Você verá que a variável "SEXOHAS" apresenta a primeira e a segunda categorias.

Em seguida, na tela da "Expressão numérica", informe o dígito "3". Clique em "Se..." para alterar a expressão. Na tela "Incluir se o caso atender à condição", informe utilizando o teclado virtual localizado no centro da janela, a expressão "SEXO = 2 & HAS = 1", ou seja, a categoria 3 = Sexo feminino hipertenso. Clique em "Continuar". Em seguida, clique em "Ok" ou em "Colar" para gerar o arquivo de sintaxe. A variável "SEXOHAS" apresenta três categorias. Por fim, para obtermos todas as categorias da nova variável "SEXOHAS", na tela "Expressão numérica", digite "4". Clique em "Se..." e, na tela "Incluir se o caso atender à condição", informe a expressão "SEXO = 2 e HAS = 0", ou seja, a categoria 4 = Sexo feminino não hipertenso. Clique em "Continuar". Em seguida, clique em "Ok" ou em "Colar" para gerar o arquivo de sintaxe. Pronto! As quatro categorias da variável "SEXOHAS" foram criadas.

Calculando variável através de outras existentes

Na opção "Calcular variável...", além de uma ampla variedade de funções matemáticas e equações complexas (como logaritmo decimal "Log10" e exponencial "Exp", localizadas na função "Aritmética") podemos calcular uma nova variável utilizando variáveis existentes no banco de dados. Como vimos até agora, o resultado de qualquer uma das funções de cálculo irá produzir uma nova coluna na janela de edição de dados. Para exemplificar, iremos criar uma variável "Índice de Massa Corporal", a partir das variáveis "Massa corporal" e "Estatura", através da fórmula: $IMC = \text{Massa corporal (em kg)} / \text{Estatura}^2 \text{ (em metros)}$.

No menu "Transformar", clique em "Calcular variável". Na caixa de diálogo "Calcular variável", clique em "Redefinir" e, em seguida, em "Variável de destino", digite o nome da nova variável, por exemplo, "IMC". Se você digitar um nome que já existe, o SPSS irá informar e perguntar "Alterar variável existente?". Se você clicar em "Ok", o SPSS substituirá os dados existentes na coluna com os resultados da função utilizada. Se você clicar em "Cancelar", você precisará renomear a nova variável. Clique em "Tipo e rótulo" e defina uma expressão para a variável, como "Índice de Massa Corporal (em kg/m²)", ou utilize a própria expressão como rótulo, clicando em "Usar expressão como rótulo". Localize na tela da esquerda a variável "Massa corporal", selecione e clique no botão de seta para copiá-la para a caixa de texto "Expressão numérica". No teclado virtual localizado no centro da janela, clique em "/" (função de divisão). Em seguida, localize na tela da esquerda a variável "Estatura", selecione e utilize o botão de seta para inseri-la na caixa de texto "Expressão numérica". No teclado virtual localizado no centro da janela, clique em "*" (função de exponenciação) e em "2". A função final será "MASSA/ESTATURA ** 2". No mesmo teclado, temos funções de adição ("+"), subtração ("-"), multiplicação ("*") e símbolos de menor que ("<"), menor que ou igual a ("<="), maior que (">"), maior que ou igual a (">="), igual a ("=") e não igual a ("!="). Clique em "Ok" ou em "Colar" para gerar o arquivo de sintaxe. Pronto! A variável "IMC" foi criada e automaticamente incluída no final do banco de dados.

Como criar uma variável a partir da contagem de outras

Muitas vezes nos deparamos com situações onde é necessário contar em quantas variáveis apareceu algum código. Por exemplo, quantas medicações o indivíduo toma, ou quantas comorbidades o paciente possui, ou ainda, quantas variáveis possuem dados omissos (*missing values*). Para isso, no comando "Valores de contagem dentro de casos", no menu "Transformar", é possível criarmos uma variável cuja informação seja uma contagem a partir de outras.

Para exemplificar, iremos contar quantos sujeitos praticam atividade física através da variável "IPAQ". Originalmente, a variável possui cinco categorias válidas: 0 = Sedentário, 1 = Insuficientemente ativo B, 2 = Insuficientemente ativo A, 3 = Ativo e, 4 = Muito ativo. No entanto, digamos que o investigador considere apenas as categorias "3" e "4" como aquelas que representam quem realmente pratica atividade física. No menu "Transformar", clique em "Valores de contagem dentro de casos...". Na janela "Ocorrência de contagem de valores dentro de casos", em "Variável de destino", digite o nome da variável que irá armazenar a contagem, como por exemplo, "PRATAF". Em "Rótulo de destino", digite uma descrição para a variável, como por exemplo "Prática de atividade física". Na tela da esquerda, selecione as variáveis (neste caso, "IPAQ") e utilize o botão da seta para inclui-las na tela "Variáveis numéricas". É possível incluir mais de uma variável na tela "Variáveis numéricas" e definir para cada uma o valor de contagem. Clique em "Definir valores...". Como definimos as categorias "3" e "4" como sendo as que definem a prática de atividade física, selecione a opção "Faixa" e informe os valores "3" até "4". Clique em "Adicionar". A faixa de valores informada será incluída na tela "Valores a contar". Clique em "Continuar". Em seguida, clique em "Ok" ou em "Colar" para gerar o arquivo de sintaxe. A nova variável "PRATAF" é exibida na janela de edição de dados. Como a variável é incluída no final do arquivo, ela será exibida na coluna à extrema direita na planilha "Visualização de Dados" e na última linha na planilha "Visualização de Variável".

Classificação de casos a partir de uma ou mais variáveis

A classificação de casos (ordenação de linhas do arquivo de dados) é frequentemente útil e, às vezes, necessária para determinados tipos de análise. Para reordenar a sequência de casos no arquivo de dados com base no valor de uma ou mais variáveis utilizamos a opção "Classificar casos" do menu "Dados". Na tela "Classificar casos", selecione a variável de interesse segundo a qual o arquivo deve ser ordenado e mova-a, através do botão da seta, para o quadro "Classificar por". Se você selecionar várias variáveis de classificação, a ordem em que elas aparecem no quadro "Classificar por" determina a ordem na qual os casos são classificados. Desta forma, o arquivo será primeiramente ordenado pelos valores da primeira variável escolhida e depois pela segunda variável. A segunda ordenação é feita para os valores comuns da primeira variável. Logo abaixo do quadro "Classificar por", existe um quadro chamado "Ordem de classificação", que permite escolher se a ordenação será "Crescente" ou "Decrescente".

Para exemplificar, faça uma ordenação segundo a variável "Sexo", selecionando uma ordem crescente, e segundo a variável "Idade", selecionando uma ordem decrescente. Mova a variável "Sexo" para o quadro "Classificar por" e escolha a opção "Crescente" no quadro "Ordem de classificação". Em seguida, mova a variável "Idade" para o quadro "Classificar por" e escolha a opção "Decrescente" no quadro "Ordem de classificação". Clique em "Ok", para executar a ordenação, ou "Colar", para gerar o arquivo de sintaxe. Atenção! Após a execução deste comando a posição dos indivíduos nas linhas fica completamente alterada, pois o indivíduo na primeira linha do banco de dados após a ordenação pode não ser o primeiro caso digitado. Para que esta informação não se perca é essencial que exista uma variável com a identificação dos casos ("ID").

Uma outra alternativa, mais prática para fazer esta ordenação, é clicar com o botão direito do mouse sobre o nome da variável na "Planilha de dados" e selecionar a ordenação ("Classificar em ordem crescente" ou "Classificar em ordem decrescente") na caixa de comando. Perceba que nesta caixa também é possível "Cortar", "Copiar", "Colar", "Eliminar" e "Inserir" uma variável.

Restringindo a análise a um subconjunto de casos

É possível restringir sua análise a um subgrupo específico com base nos critérios que incluem variáveis e expressões complexas ou também selecionar uma amostra aleatória de casos. Os critérios usados para definir um subgrupo podem incluir valores de variáveis e intervalos, intervalos de data e hora, números de casos (linhas), expressões aritméticas e funções. Essa seleção de parte do arquivo de dados pode ser uma necessidade temporária, como quando queremos verificar alguma condição em determinado grupo, ou permanente, como quando estamos interessados em estudar apenas um segmento específico da amostra. O SPSS possui várias formas de seleção de dados. Para fazer qualquer tipo de seleção, utilizamos a opção "Selecionar casos..." no menu "Dados".

Na janela "Selecionar casos", estão presentes cinco opções diferentes para seleção, descritas no quadro "Selecionar":

1. "Todos os casos": opção utilizada por padrão e inclui todas as observações do banco de dados, ou seja, nenhuma seleção é executada. Clique em "Ok" para selecionar todos os caso, ou em "Colar" para gerar o arquivo de sintaxe.

2. "Se a condição for cumprida": através dessa opção, podemos definir expressões condicionais para seleção de casos, inclusive expressões aritméticas e funções. Para isso, indicamos a condição clicando em "Se...". Na caixa de diálogo "Selecionar casos: Se", a expressão condicional pode usar nomes de variáveis existentes, constantes, operadores aritméticos, operadores lógicos, operadores relacionais e funções. É possível digitar e editar a expressão na caixa de texto apenas como texto em uma janela de saída. É possível também usar o teclado da calculadora, a lista de variáveis e a lista de funções para colar elementos na expressão. Desta forma, você pode mover para o quadro superior as variáveis que estão na tela da esquerda, bem como selecionar as funções da tela "Grupo de função" e inserir uma função específica da tela "Funções e variáveis

especiais” no mesmo quadro. Por exemplo, para avaliar apenas os sujeitos ativos (“3”) e muito ativos (“4”) da variável “Classificação IPAQ”, selecione a variável “IPAQ” na tela da esquerda, insira-a no quadro superior com o botão da seta e informe, utilizando o teclado virtual localizado no centro da janela, a primeira parte da expressão: “IPAQ = 3”. Para acrescentar a outra categoria da mesma variável “IPAQ”, insira o símbolo “|” e adicione o restante da expressão: “IPAQ = 4”. Após esse procedimento, o quadro superior deve conter a seguinte expressão: “IPAQ = 3 | IPAQ = 4”, ou seja, sujeitos ativos (“3”) e muito ativos (“4”) da variável “IPAQ” serão selecionados. Nota: utilize o símbolo “&” quando forem selecionadas categorias de diferentes variáveis. Clique em “Continuar”. Em seguida, clique em “Ok” ou em “Colar” para gerar o arquivo de sintaxe. Na janela de edição de dados, na planilha “Visualização de dados”, você perceberá algumas alterações após a seleção: os casos (linhas) que não foram selecionadas apresentam-se tachadas com uma linha diagonal no meio do número da linha; a barra localizada na parte inferior da janela apresenta a mensagem “Filtro Ativado”; além disso, uma coluna de nome “filter_\$” é adicionada. Essa nova coluna apresenta valor “0” para as linhas que não foram selecionadas e valor “1” para as linhas que foram selecionadas.

3. “Amostra aleatória de casos”: podemos selecionar uma porcentagem ou número exato de casos. A seleção é feita aleatoriamente. Para isso, clicamos em “Amostra”. Na caixa de diálogo “Selecionar casos: Amostra aleatória”, é possível selecionar uma das seguintes alternativas para o tamanho da amostra: a) “Aproximadamente”: em que será selecionada uma porcentagem especificada pelo investigador, gerando uma amostra aleatória aproximadamente da porcentagem especificada de casos; b) “Exatamente”: neste caso, um número de casos especificado pelo usuário será selecionado. Você deve também especificar o número de casos a partir dos quais deve ser gerada a amostra. Este segundo número deve ser menor ou igual ao número total de casos no arquivo de

dados. Se o número exceder o número total de casos no arquivo de dados, a amostra conterá proporcionalmente menos casos do que o número solicitado.

4. "Com base na hora ou faixa de casos": usamos essa opção quando estamos interessados em selecionar uma faixa específica de valores, por exemplo, os casos do número 50 ao 100, ou também podemos fazer seleções baseadas em datas. Para isso, informe o intervalo em "Faixa...". Na caixa de diálogo "Selecionar casos: Faixa", é possível selecionar um intervalo de números de observações (linhas). Na "Observação", no "Primeiro caso", insira o número de observação inicial (número da linha na janela de edição de dados). Se você não especificar um valor de "Último caso", todos os casos a partir do número de observação inicial serão selecionados. No "Último caso", insira o número de observação final (número da linha na janela de edição de dados). Se você não especificar um valor de "Primeiro Caso", todos os casos a partir do número de observação final serão selecionados.

5. "Usar variável de filtro": nesta opção, uma variável é escolhida no banco de dados e usada como filtro. Todos os casos para os quais a variável filtro assume o valor "0" não serão selecionados.

Na mesma janela, no quadro "Saída", você pode:

1. "Não incluir na filtragem os casos não selecionados": os casos (linhas) que não são selecionados não são incluídos nas análises posteriores, porém, permanecem na janela de dados. Caso você queira usar os casos não selecionados, na mesma janela, no quadro "Selecionar", opte por incluir "Todos os casos" na seleção. Se você selecionar uma amostra aleatória ou se selecionar casos com base em uma expressão condicional, isso gerará uma variável denominada "filter_\$" com um valor de "1" para casos selecionados e um valor de "0" para casos não selecionados;

2. "Copiar casos selecionados para um novo conjunto de dados": nesta opção, o SPSS cria uma nova janela de dados

com os casos selecionados, ou seja, os casos selecionados são copiados para um novo conjunto de dados, deixando o conjunto de dados original não afetado. Os casos não selecionados não são incluídos no novo conjunto de dados e são deixados em seu estado original no conjunto de dados original. O nome da nova janela de dados deve ser fornecido pelo usuário na lacuna "Nome do conjunto de dados".

3. "Excluir os casos não selecionados": os casos (linhas) não selecionados são apagados (excluídos) da janela de dados. Os casos excluídos podem ser recuperados apenas saindo do arquivo sem salvar nenhuma mudança e, em seguida, reabrindo o arquivo. A exclusão de casos será permanente se você salvar as mudanças no arquivo de dados. Neste caso deve-se tomar o cuidado de salvar o banco de dados com outro nome, em "Analisar", "Salvar como...".

Lembre-se: para retornar todo o conjunto de dados, basta selecionar a opção "Todos os casos", no quadro "Selecionar".

Análises simultâneas em diferentes subconjuntos

Em algumas situações, o pesquisador pode querer analisar diferentes grupos de uma amostra de forma idêntica, mas em particular para cada grupo. Por exemplo, vamos supor que o pesquisador queira analisar separadamente as medidas pressóricas para homens e mulheres, pois verificou que essas variáveis se comportam de forma completamente diferente entre os sexos. Não faz sentido, portanto, apresentar a análise destes dados considerando ambos os sexos. Ou seja, o que pretendemos fazer, daqui para frente, são duas análises idênticas, uma para cada sexo.

A opção "Dividir arquivo...", no menu "Dados", permite repetir qualquer análise sobre vários grupos de casos. Essa função permite especificar uma variável de grupo (lembre-se que essas variáveis são utilizadas para especificar categorias da unidade experimental). Qualquer procedimento posterior no SPSS será executado em cada uma das categorias especificadas pela variável de grupo.

Por padrão, o SPSS considera todos os casos juntos na análise, sem separação por grupos. Por esse motivo, a opção selecionada na janela "Dividir arquivo" é "Analisar todos os casos, não criar grupos". Para executar a análise para as categorias de uma determinada variável, podemos optar por "Comparar grupos" ou "Organizar saída por grupos" e, então, o quadro "Grupos baseados em" fica disponível. Movemos para o quadro "Grupos baseados em" a(s) variável(is) que definirá(ão) os grupos para os quais a análise deve ser repetida. Se mais do que uma variável for selecionada, os grupos serão definidos pela combinação das categorias de todas as variáveis. Logo abaixo, podemos ainda escolher se o banco de dados deve ser ordenado pela variável que definirá os grupos ("Classificar o arquivo agrupando variáveis") ou se o banco de dados já está ordenado pela variável que definirá os grupos ("O arquivo já foi classificado"). Se o banco não estiver ordenado corretamente, o comando não será executado de forma adequada.

Se você selecionar "Comparar grupos", os resultados de todos os grupos de divisão de arquivo serão incluídos na mesma tabela. Se você selecionar "Organizar saída por grupos", duas tabelas dinâmicas serão criadas, uma para cada grupo. Se o processamento de divisão de arquivo estiver ativo, a mensagem "Dividir por" aparecerá na barra de status na parte inferior da janela de edição dos dados.

Para exemplificar, clique em "Dividir arquivo...", no menu "Dados", marque a opção "Comparar grupos", selecione na tela da esquerda a variável "Sexo" e mova-a com o botão da seta para o quadro "Grupos baseados em". Em seguida, clique em "Ok" ou em "Colar" para gerar o arquivo de sintaxe. Lembre-se: após a execução desse comando, todas as análises feitas gerarão dois resultados, um para o sexo masculino e outro para o sexo feminino.

Note que a única mudança observada na janela de edição de dados é a mensagem "Dividir por SEXO" na barra inferior direita, ou a ordenação dos casos pela variável que definiu os grupos, caso o banco de dados ainda não estivesse ordenado. Faça uma análise descritiva para a variável "Colesterol total", por exemplo, para verificar como os resultados serão apresentados no arquivo de saída. Para isso, clique em "Analisar", "Estatísticas descritivas..."

e “Frequências...”. Na janela “Frequências”, selecione na tela da esquerda a variável “Colesterol total” e mova-a para a lista de “Variável(is)”. Clique em “Estatísticas” e solicite medidas de tendência central e de dispersão. Em seguida, clique em “Ok” ou em “Colar” para gerar o arquivo de sintaxe. Veja que os resultados são apresentados em um único bloco, mas separando em primeiro o sexo masculino e em segundo o sexo feminino.

Atenção! Quando você tiver terminado com o comando de divisão de arquivo, lembre-se de desligá-lo. Retorne à caixa de diálogo “Dividir arquivo” e selecione a opção “Analise todos os casos, não crie grupos”.

Inclusão de novos casos ou novas variáveis

Muitas vezes os dados se encontram em diferentes bancos de dados e em algum momento precisamos uni-los para poder cruzar alguma variável que está num banco com outra de outro banco ou ainda unificar em um mesmo banco, casos de diferentes bancos de dados com as mesmas variáveis. O SPSS fornece uma opção para incluir um banco de dados dentro de outro sem que seja necessário digitar todas as variáveis novamente. Nesse caso, os dois bancos precisam estar no formato SPSS, ou seja, no formato “.sav”. Esse procedimento é feito pela opção “Mesclar arquivos” no menu “Dados”, que permite a união de bancos de dados que tenham 1) as mesmas variáveis com diferentes sujeitos (“Adicionar casos”) ou 2) variáveis diferentes dos mesmos sujeitos (“Adicionar variáveis”).

Para a primeira opção, em “Dados”, selecione “Adicionar casos” em “Mesclar arquivos”. Neste caso, é imprescindível que os dois bancos de dados tenham as mesmas variáveis, com o mesmo nome e o mesmo formato. Para exemplificar, vamos juntar o “Banco de dados 3.sav” e o “Banco de dados 3 casos novos.sav” (disponível em <https://bit.ly/bancosdedados>), que possuem diferentes indivíduos cada um, mas o mesmo conjunto de variáveis. Na janela “Adicionar casos”, é possível escolher uma das seguintes opções: 1) “Um conjunto de dados aberto”: caso o outro banco de dados (“Banco de dados 3 casos novos.sav”) já

esteja aberto no SPSS, ou 2) "Um arquivo de dados externo do SPSS *Statistics*": se o banco de dados não estiver aberto. Clique em "Navegar" para localizar o arquivo, selecione o banco de dados e clique em "Continuar". Após selecionado o banco de dados, na caixa de diálogo seguinte, serão listadas as "Variáveis sem par". Isso significa que existem variáveis no primeiro banco de dados que não constam no segundo banco de dados. As variáveis seguidas de "*" estarão no primeiro banco de dados e aquelas seguidas de "+" estarão no segundo banco de dados e ambas não serão incluídas no novo banco de dados. Se você deseja utilizar as variáveis sem par em análises futuras, é recomendado que você salve o banco de dados original antes de adicionar os novos casos. As variáveis comuns serão automaticamente incluídas na tela "Variáveis no novo conjunto de dados ativo". Se você quiser identificar qual o banco de origem de cada caso no novo banco, clique em "Indicar origem de caso como variável". Ao clicar em "Ok" ou em "Colar", o comando será executado e os novos casos serão adicionados ao final do banco de dados. Se você solicitou "Indicar origem de caso como variável", haverá uma coluna indicando casos antigos ("0") e novos ("1").

Na segunda opção, em "Dados", selecione "Adicionar variáveis" em "Mesclar arquivos". Neste caso, é imprescindível que os dois bancos de dados não possuam as mesmas variáveis, mas possuam os mesmos sujeitos e que os casos estejam ordenados. Para exemplificar, vamos juntar as variáveis do "Banco de dados 3.sav" e do "Banco de dados 3 questionário.sav" (disponível em <https://bit.ly/bancosdedados>). Neste último, temos variáveis de um questionário de sintomas neuropáticos. Primeiro, precisamos nos certificar que os casos estejam ordenados em ambos os bancos de dados. Para isso, como vimos neste capítulo é possível ordenar os casos clicando com o botão direito do mouse sob a coluna de identificação dos casos (variável "ID") e selecionar "Classificar em ordem crescente". Da mesma forma vista anteriormente, é possível escolher, na janela "Adicionar variáveis": 1) "Um conjunto de dados aberto": caso o outro banco de dados ("Banco de dados 3 questionário.sav") já esteja aberto no SPSS, ou 2) "Um arquivo de dados externo do SPSS *Statistics*": se o banco de dados não estiver aberto. Clique em "Navegar" para localizar o

arquivo, selecione o banco de dados e clique em "Continuar". Na caixa de diálogo seguinte, selecione a opção "Combinar casos em variáveis chave em arquivos classificados", e opte por uma das seguintes opções: 1) "Ambos os arquivos fornecem casos": se cada banco possui uma linha para cada indivíduo da amostra, 2) "O conjunto de dados não ativo é uma tabela vinculada" ou 3) "O conjunto de dados ativo é uma tabela vinculada": se um dos bancos possuir mais de um registro (linha) por caso, estes devem estar identificados pela mesma variável chave. Opte pela primeira opção e selecione a variável "ID" que identifica o número dos casos (deve estar na tela "Variáveis excluídas"). Clique no botão da seta para inseri-la na caixa "Variáveis-chave". Na janela "Novo conjunto de dados ativo", as variáveis seguidas de "*" estão no primeiro banco de dados e aquelas seguidas de "+" estão no segundo banco de dados e correspondem as variáveis que serão incluídas. Em seguida, clique em "Ok" ou em "Colar". Uma mensagem de advertência aparecerá: "A correspondência vinculada falhará se os dados não forem classificados na ordem crescente de variáveis-chave". Clique em "Ok". Note que as variáveis do questionário serão automaticamente incluídas no final do banco de dados.

Referências

Field A. Descobrimo a estatística usando o SPSS. Tradução: Lorí Viali. 2. ed. Porto Alegre: Artmed, 2009. 684 p.

Exercícios sugeridos

1. Utilizando o "Banco de dados 3.sav":

a) Categorize a variável "IMC", criada neste capítulo, em graus de obesidade, onde: Obesidade grau I = 30,00 a 34,99 kg/m²; Obesidade grau II = 35,00 a 39,99 kg/m² e; Obesidade grau III = >40,0 kg/m².

b) Crie uma variável de Pressão Arterial Média (PAM), através das medidas pressóricas, utilizando a expressão: $PAM = ((2 * PAD) + PAS) / 3$. Onde: PAS = Pressão Arterial Sistólica e PAD = Pressão Arterial Diastólica.

c) Crie uma variável de LDL-Colesterol (LDL), utilizando a Fórmula de Friedewald: $LDL = CT - HDL - (TG / 5)$. Onde: CT = Colesterol total, HDL = HDL-Colesterol e TG = Triglicerídeos.

d) Após categorizar e calcular as variáveis, faça uma análise descritiva das variáveis pressóricas (PAS, PAD e PAM) e lipídicas (Colesterol total, Triglicerídeos, HDL-Colesterol e LDL-Colesterol), selecionando somente os indivíduos com obesidade grau III. Construa uma tabela apresentando os resultados e interprete-os supondo uma distribuição normal dos dados.

e) Faça a mesma descritiva separando por sexo, sem discriminar o grau de obesidade (incluindo todos os casos). Construa uma tabela apresentando os resultados e interprete-os supondo uma distribuição normal dos dados.

2. Utilizando o "Banco de dados 1.sav", inclua os novos casos do "Banco de dados 1 casos novos.sav".

3. Em seguida, utilizando o "Banco de dados 1.sav" com os novos casos incluídos, divida o banco de dados pelo nível de pós-graduação e responda:

a) Qual a média de idade do nível mestrado e do nível doutorado?

b) Qual a frequência de sexo feminino no nível mestrado e no nível doutorado?

c) Em que período se encontram a maioria dos estudantes do nível mestrado e do nível doutorado?