Capítulo 4

Criação de Bancos de Dados

Ângelo Giuseppe Roncalli da Costa Oliveira

INTRODUÇÃO

As pesquisas epidemiológicas, assim como todas as pesquisas de caráter quantitativo, pressupõem uma seqüência de etapas que vão do planejamento da pesquisa até a elaboração do relatório final, passando pela coleta e processamento dos dados. Uma atribuição precípua da estatística aplicada aos estudos epidemiológicos é a consolidação de dados obtidos de amostras ou populações de modo que estes possam ser lidos e interpretados em seu conjunto.

Desse modo, a etapa subseqüente à coleta de dados é a construção de um banco em que tais dados sejam organizados de forma a facilitar as tarefas de análise. Este capítulo abordará as estratégias necessárias para a construção de bancos de dados em pesquisas epidemiológicas, dando especial destaque à tabulação eletrônica a partir dos principais programas de gerenciamento de dados.

ENTENDENDO AS VARIÁVEIS

As pesquisas epidemiológicas envolvem o estudo de características que não são distribuídas de modo uniforme na população. O conceito de "variável" refere-se justamente a essas características populacionais não-uniformes, que se propõe descrever e analisar no âmbito das pesquisas epidemiológicas. *Grosso modo*, "variável" pode ser definida como a expressão numérica de qualquer evento da natureza. É tudo aquilo que se deseja estudar e que pode ser traduzido em números, seja através de contagem, mensuração ou classificação. As variáveis, portanto, estão associadas a eventos contábeis, mensuráveis ou classificáveis; e, considerando a natureza complexa dos objetos de estudo da epidemiologia, possuem limitações diretamente proporcionais à subjetividade do evento. Ao contarmos uma certa quantidade de eventos ou medirmos alguns deles, geramos variáveis ditas *quantitativas*; ao classificarmos os eventos, obtemos variáveis do tipo *categóricas*. Peso, altura, CPO, glicemia são exemplos de variáveis quantitativas, e sexo, etnia, grau de instrução e moradia são exemplos de variáveis categóricas (Berquó, 1981).

PRINCÍPIOS GERAIS PARA A CRIAÇÃO DE BANCOS DE DADOS

Com os recursos tecnológicos atualmente disponíveis, não se admite mais que os dados envolvidos em pesquisas epidemiológicas sejam tabulados manualmente. Além de demorada, desgastante e limitada, a tabulação manual submete o estudo a um risco elevado de erros. Com o advento e a disseminação da informática, a tabulação eletrônica tornou a análise de dados epidemiológicos muito mais rápida, eficiente e segura. Com isso, a descrição e a análise dependem, fundamentalmente, de uma cuidadosa elaboração do banco de dados da pesquisa. A correspondência entre o banco de dados e o instrumento da coleta de dados na pesquisa facilita a digitação e, posteriormente, a análise dos dados. Outras recomendações importantes são:

1. Estabelecimento de códigos para as variáveis categóricas

A codificação das variáveis pode ser efetuada durante a construção do instrumento de coleta de dados. Caso isso não tenha sido feito, a codificação poderá ser realizada por ocasião da entrada dos dados em uma base eletrônica. Os códigos devem ser, preferencialmente, numéricos e com um único dígito, a não ser, obviamente, quando se trabalha com variáveis quantitativas que demandam outras escalas de medida. Podem ser usadas letras como códigos, quando o número de categorias passa de 10, ou quando é necessário estabelecer uma diferenciação entre as categorias, como é o caso dos registros de condição dentária, em que os códigos numéricos dizem respeito aos dentes permanentes e as letras correspon-

dem aos dentes decíduos. A utilização de códigos numéricos facilita bastante a digitação, pelo fato de permitirem efetuar a digitação exclusivamente através do teclado numérico do computador, uma estratégia muito utilizada por digitadores profissionais.

2. Criação de códigos de exclusão

Embora alguns programas de bancos de dados ignorem as células deixadas em branco na análise, recomenda-se evitar deixar a variável sem preenchimento, para evitar confusão. A maioria dos programas permite que um determinado código, por exemplo o algarismo 9, seja interpretado como informação não-disponível (*missing*), o que facilita bastante a análise.

3. Utilização de dados quantitativos brutos

Na medida do possível, o dado deve ser captado em sua expressão numérica primária, evitando categorias estabelecidas *a priori*. Essa recomendação é útil tanto na construção do instrumento de coleta de dados como na criação do banco informatizado. Ao se avaliar a renda mensal familiar, por exemplo, é mais prático captar a renda em reais para somente durante a análise estabelecer as faixas de renda ou transformação em outra unidade, como salários mínimos. Ao se obter a informação já incluída em faixas preestabelecidas, perde-se a informação original, além de haver o risco de uma distribuição heterogênea da variável entre os elementos amostrais. A classificação de faixas de renda (por exemplo, menos de 1 salário mínimo, de 1 a 2 e 2 ou mais salários mínimos) pode ser muito útil para pesquisas envolvendo população de baixa renda, mas teria pouca utilidade em bairros de classe média alta. Outro exemplo diz respeito à escolaridade, que pode ser expressa em número de anos de estudo, evitando a obtenção da informação por graus (ensino fundamental, médio e superior).

4. Critérios de validação de entrada

Os programas de bancos de dados permitem a criação de critérios de validação de entrada de dados. Isso é particularmente importante quando diferentes digitadores contribuem para a informatização dos dados e diminui consideravelmente os erros de digitação.

5. Verificação de erros de digitação

O cumprimento da recomendação anterior reduz o risco de erros de digitação. Mesmo assim, podem ocorrer erros quando se digitam dados válidos porém não correspondentes ao registro que consta na ficha de coleta. Em alguns casos, recomenda-se a digitação dupla ou mesmo tripla para minimizar o risco de erros. Após o banco pronto, ainda deve ser realizada uma avaliação, por amostragem, do percentual de erros de digitação. A simples verificação da distribuição de freqüência das variáveis em estudo possibilita a identificação de valores aberrantes, possivelmente fruto de erros de digitação ou anotação, permitindo assim sua correção.

6. Criação de página de códigos

Em função das recomendações anteriores, é importante criar uma tabela em que sejam explicitadas as informações relativas ao banco de dados, particularmente os códigos empregados. Em alguns programas, como o SPSS ou o Microsoft Excel, essa informação faz parte da estrutura do banco. Quando for necessário disponibilizar o banco de dados em uma linguagem de uso comum para diferentes programas de informática, como os arquivos de extensão DBF, é necessário apresentar em anexo a tabela de códigos correspondente, como exemplificado no Quadro 4.1.

CRIANDO UM BANCO DE DADOS

Há uma quantidade considerável de programas de bancos de dados. Alguns são mais sofisticados e exigem conhecimentos de programação, sendo mais aplicáveis às áreas comercial e financeira. Especificamente para pesquisas epidemiológicas, existem bons programas que permitem a construção do banco de dados e sua posterior análise. A despeito de cada um deles possuir suas especificidades, a lógica de criação dos bancos de dados é muito semelhante entre eles, bastando, na maioria dos casos, seguir as recomendações anteriores. Descreveremos aqui as informações mais importantes para a criação de bancos em três dos mais populares programas disponíveis, o Microsoft Excel®, o SPSS (Statistical Package for Social Science) e o Epi Info.

Utilizando o Microsoft Excel®

O Microsoft Excel® é uma conhecida planilha eletrônica integrada ao pacote de aplicativos de "escritório" mais utilizado em computadores pessoais no Brasil, o Microsoft Office. Na verdade, o Excel não é um programa de banco de dados e, em princípio, poderia não ser o aplicativo mais adequado para se trabalhar com dados epidemiológicos. Contudo, a facilidade de seu uso, sua versatilidade e popularidade permitem a construção de bancos de dados relativamente simples, quando se trabalha com dados numéricos e/ou categóricos. Para questionários mais complexos, com campos descritivos, recomenda-se utilizar programas específicos para questionários, como o Epi Info ou, caso se tenha em mente alguma análise de caráter qualitativo, há outras opções como o Evoc ou Alcest. Mas, para estudos envolvendo o cálculo de medidas de tendência central e de dispersão, bem como freqüências absolutas e percentuais, o Excel é uma boa opção, embora não contemple recursos para análises de inferência.

Outra vantagem do Excel é que o formato XLS de seus arquivos é passível de leitura direta por boa parte dos progra-

QUADRO 4.1 Descrição das variáveis constantes em um banco de dados

Variável	Descrição	Tipo	Categorias
UF	Unidade da Federação	Categórica Nominal	Código do IBGE
FLUOR	Presença de água fluoretada	Categórica Nominal	1- Fluoretado 2- Não-fluoretado 9- Sem Informação
ESCOLA	Tipo de escola	Categórica Nominal	1- Pública 2- Privada 9- Sem Informação
IDADE	Idade em anos	Quantitativa Discreta	Dado numérico
SEXO	Sexo	Categórica Nominal	1- Masculino 2- Feminino 9- Sem Informação
DENTAL16	Código CPO para dente 16	Categórica Nominal	0- Hígido 1- Cariado 2- Restaurado com cárie 3- Restaurado sem cárie 4- Extraído por cárie 5- Extraído por outras razões 9- Não-examinado

Fonte: Exemplo retirado do banco de dados do Projeto SBBrasil 2003 (Brasil, 2004).

mas estatísticos, como o SPSS, o S+ ou o Statistica®. Além disso, pode exportar para outros formatos, como o Dbase (DBF – Data Base File), um padrão quase universal de bancos de dados. Ademais, em geral, seus arquivos não são muito grandes (a não ser quando se exagera nas formatações de cores e linhas), o que facilita a troca de informações em meio virtual. O Excel tem uma capacidade razoável de armazenamento de dados e serve bem a muitas finalidades das pesquisas epidemiológicas. O limite de linhas em uma planilha é de 65.536, o que significa que pesquisas com um número de unidades amostrais superior a esse limite terão que recorrer a outros programas. Sua capacidade de armazenamento para variáveis é bastante considerável, permitindo o uso de até 256 colunas.

ETAPAS PARA A CONSTRUÇÃO DE BANCOS NO EXCEL

Considera-se que o leitor possua conhecimentos básicos sobre o funcionamento de planilhas eletrônicas,* como as operações comuns aos programas do pacote Microsoft Office, tais como salvamento de arquivos, impressão, formatação, entre outras. Considerações mais complexas, relativas à análise de dados e à construção de tabelas e gráficos, não serão tratadas neste capítulo.

*Foi utilizada, para os exemplos deste capítulo, a versão 2002 para Windows XP. Versões anteriores para Windows 98 ou da família do Office 2000 podem apresentar ligeiras diferenças na apresentação visual e nas funções.

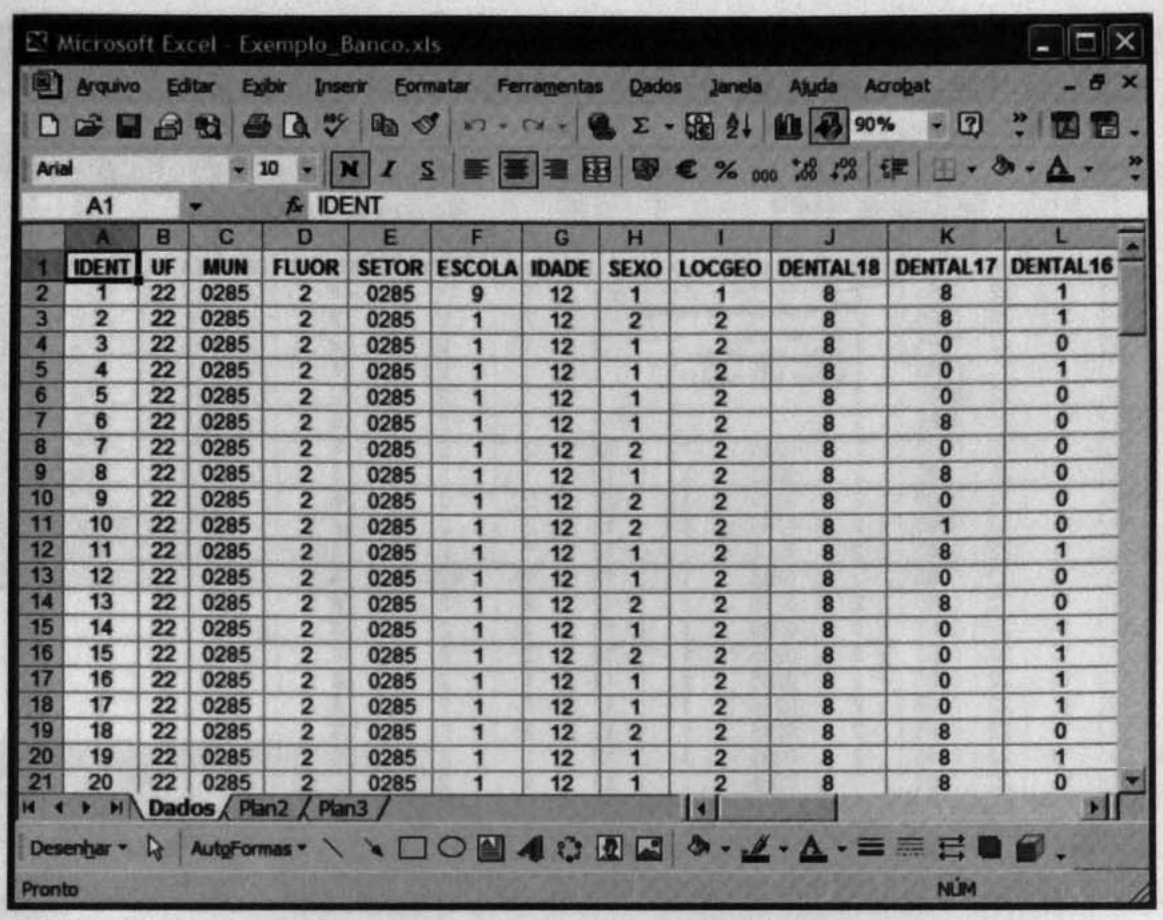
1. Definição das variáveis

Em primeiro lugar, devem ser definidas as variáveis que constituirão o banco de dados, com suas respectivas codificações, conforme indicado anteriormente. Essas variáveis preencherão toda a primeira linha da planilha, sendo cada linha subseqüente reservada para o preenchimento das informações relativas a cada elemento da amostra.

É importante observar que, embora o programa aceite nomes com tamanho ilimitado, recomendam-se algumas precauções ao nomear as variáveis. Entre essas precauções, sugere-se não exceder oito caracteres, bem como evitar o uso de cedilha, acentos, traços (a não ser o traço subscrito ou "*underline*") e espaços. Isso se justifica para facilitar o processo de exportação do arquivo para outras plataformas, as quais solicitam essas restrições. O SPSS, por exemplo, apenas aceita os primeiros oitos caracteres registrados para os nomes de variáveis, e o Epi Info aceita até 10 caracteres. Do mesmo modo, ao se utilizar o Epi Info em sua versão 6.04 em MS-DOS,* somente caracteres padrão serão aceitos. Além disso, ao serem criadas variáveis com nomes de extensão reduzida, o banco fica menor e mais fácil de ser utilizado.

O banco de dados apresentado como exemplo, a seguir, é o produto de um formato de pesquisa muito comum em epidemiologia em saúde bucal, o levantamento epidemiológico. As variáveis que constam no arquivo advêm da ficha básica

*O MS-DOS (Microsoft Disk Operating System) foi um dos primeiros sistemas operacionais a serem utilizados em computadores pessoais. Tinha uma interface ainda pouco amigável e hoje está praticamente em desuso com o advento do sistema Windows.



Microsoft Excel - Exemplo_Banco.xls

Arquivo Editar Exibir Inserir Formatar Ferramentas Dados Janela Ajuda Acrobat

Arial 10 N I S 90%

A1 IDENT

	A	B	C	D	E	F	G	H	I	J	K	L
1	IDENT	UF	MUN	FLUOR	SETOR	ESCOLA	IDADE	SEXO	LOCGEO	DENTAL18	DENTAL17	DENTAL16
2	1	22	0285	2	0285	9	12	1	1	8	8	1
3	2	22	0285	2	0285	1	12	2	2	8	8	1
4	3	22	0285	2	0285	1	12	1	2	8	0	0
5	4	22	0285	2	0285	1	12	1	2	8	0	1
6	5	22	0285	2	0285	1	12	1	2	8	0	0
7	6	22	0285	2	0285	1	12	1	2	8	8	0
8	7	22	0285	2	0285	1	12	2	2	8	0	0
9	8	22	0285	2	0285	1	12	1	2	8	8	0
10	9	22	0285	2	0285	1	12	2	2	8	0	0
11	10	22	0285	2	0285	1	12	2	2	8	1	0
12	11	22	0285	2	0285	1	12	1	2	8	8	1
13	12	22	0285	2	0285	1	12	1	2	8	0	0
14	13	22	0285	2	0285	1	12	2	2	8	8	0
15	14	22	0285	2	0285	1	12	1	2	8	0	1
16	15	22	0285	2	0285	1	12	2	2	8	0	1
17	16	22	0285	2	0285	1	12	1	2	8	0	1
18	17	22	0285	2	0285	1	12	1	2	8	0	1
19	18	22	0285	2	0285	1	12	2	2	8	8	0
20	19	22	0285	2	0285	1	12	1	2	8	8	1
21	20	22	0285	2	0285	1	12	1	2	8	8	0

Dados / Plan2 / Plan3

Desenhar AutoFormas

Pronto NUM

FIG. 4.1 Exemplo de banco de dados no Excel.

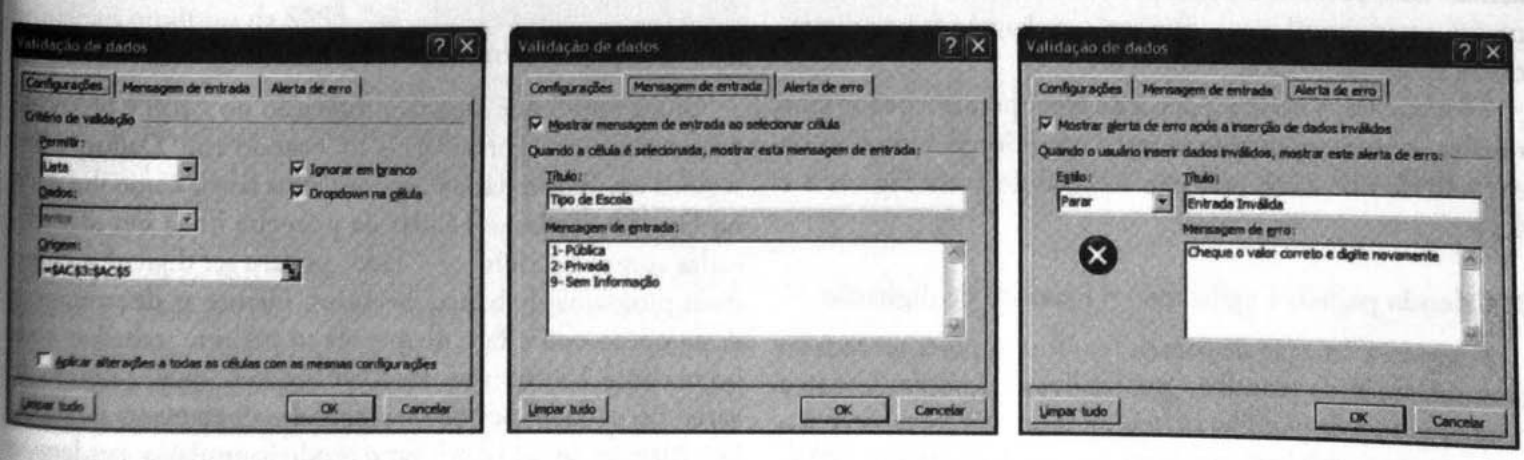
proposta pela OMS e adaptada recentemente para o Projeto SBBrasil 2003 (vide o Cap. 3 da Parte I). Observe que, na planilha, foram incluídas as variáveis na primeira linha com o cuidado de criar uma variável IDENT (Identificação), onde os elementos amostrais são devidamente numerados.

2. Validando a entrada de dados

Após criar a estrutura do banco de dados e antes de começar a digitação, é importante acrescentar alguns aperfeiçoamentos disponibilizados pelas ferramentas do Excel. Em primeiro lugar, como em qualquer banco de dados, podem ser criadas regras de validação para a entrada dos dados, o que agiliza o processo e evita erros de digitação.

Para criar essas regras, marca-se a coluna da variável, clica-se em "Dados" e escolhe-se a opção "Validação". Em seguida abre-se uma caixa de diálogo como a ilustrada na Fig. 4.2, que exemplifica a validação da variável "Tipo de Escola".

Na primeira interface de diálogo que se abre (usualmente denominada "orelha"), o item "Configurações" permite que se informe que códigos podem ser aceitos para aquela variá-



Validação de dados

Configurações | Mensagem de entrada | Alerta de erro

Critério de validação

Permitir: Lista — Ignorar em branco — Dropdown na célula

Dados:

Origem: =$AC$3:$AC$5

Aplicar alterações a todas as células com as mesmas configurações

Limpar tudo OK Cancelar

Validação de dados

Configurações | Mensagem de entrada | Alerta de erro

Mostrar mensagem de entrada ao selecionar célula

Quando a célula é selecionada, mostrar esta mensagem de entrada:

Título: Tipo de Escola

Mensagem de entrada:

1- Pública
2- Privada
9- Sem Informação

Limpar tudo OK Cancelar

Validação de dados

Configurações | Mensagem de entrada | Alerta de erro

Mostrar alerta de erro após a inserção de dados inválidos

Quando o usuário inserir dados inválidos, mostrar este alerta de erro:

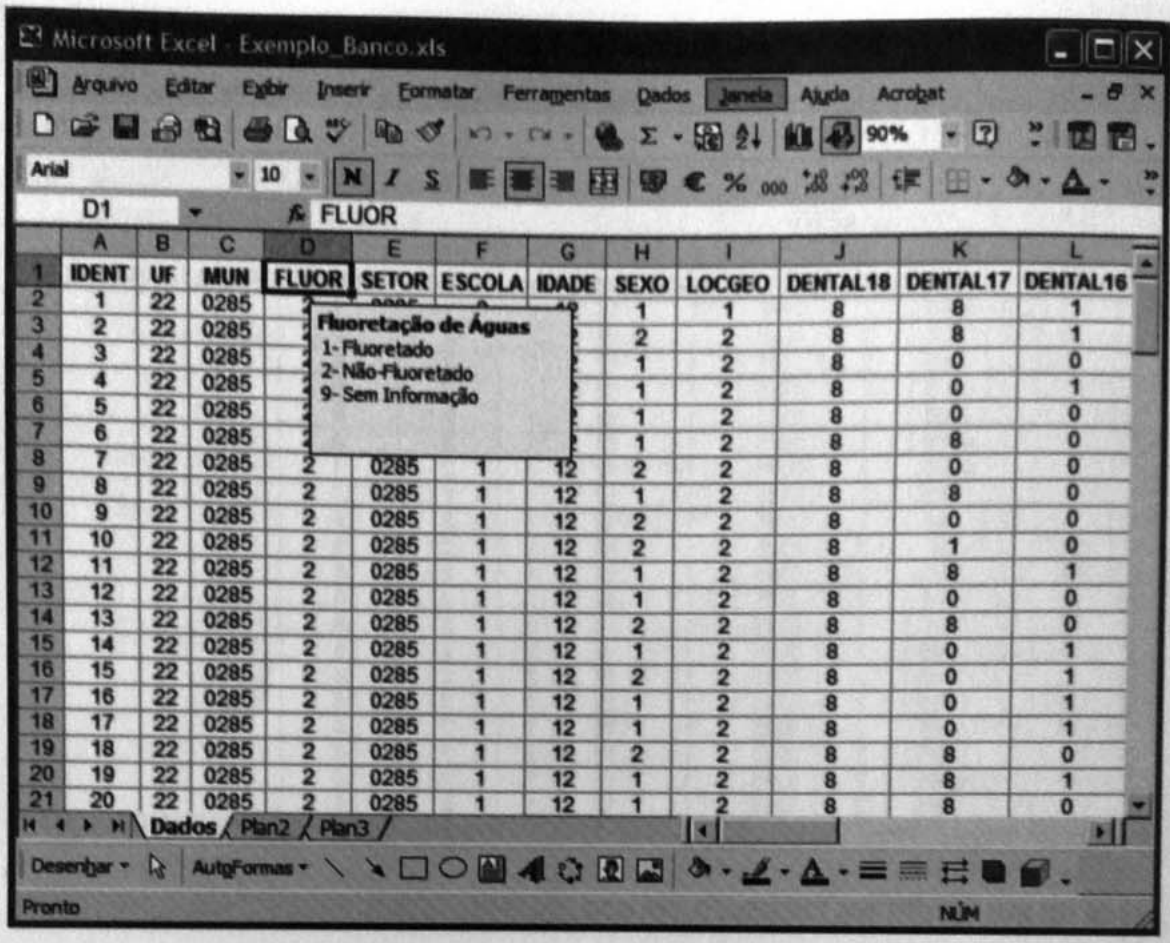
Estilo: Parar

Título: Entrada Inválida

Mensagem de erro: Cheque o valor correto e digite novamente

Limpar tudo OK Cancelar

FIG. 4.2 Caixa de diálogo das funções para validação da entrada de dados no Excel.



	A	B	C	D	E	F	G	H	I	J	K	L
1	IDENT	UF	MUN	FLUOR	SETOR	ESCOLA	IDADE	SEXO	LOCGEO	DENTAL18	DENTAL17	DENTAL16
2	1	22	0285					1	1	8	8	1
3	2	22	0285					2	2	8	8	1
4	3	22	0285					1	2	8	0	0
5	4	22	0285					1	2	8	0	1
6	5	22	0285					1	2	8	0	0
7	6	22	0285					1	2	8	8	0
8	7	22	0285	2	0285	1	12	2	2	8	0	0
9	8	22	0285	2	0285	1	12	1	2	8	8	0
10	9	22	0285	2	0285	1	12	2	2	8	0	0
11	10	22	0285	2	0285	1	12	2	2	8	1	0
12	11	22	0285	2	0285	1	12	1	2	8	8	1
13	12	22	0285	2	0285	1	12	1	2	8	0	0
14	13	22	0285	2	0285	1	12	2	2	8	8	0
15	14	22	0285	2	0285	1	12	1	2	8	0	1
16	15	22	0285	2	0285	1	12	2	2	8	0	1
17	16	22	0285	2	0285	1	12	1	2	8	0	1
18	17	22	0285	2	0285	1	12	1	2	8	0	1
19	18	22	0285	2	0285	1	12	2	2	8	8	0
20	19	22	0285	2	0285	1	12	1	2	8	8	1
21	20	22	0285	2	0285	1	12	1	2	8	8	0

FIG. 4.3 Aspecto da planilha após a validação, quando a célula é selecionada.

vel. Há diversas opções para se realizar essa operação, como escolher uma lista, estabelecer um intervalo numérico, entre outras. Para o caso em que se trabalha com listas de códigos, deve-se colocar essa lista de códigos em outro ponto da planilha e informar, na caixa de diálogo, onde se encontra essa informação.

Na orelha "Mensagem de entrada" pode-se optar pela visualização de uma mensagem de ajuda informando os códigos válidos, que aparece quando o cursor passa pela célula. Esse recurso é particularmente útil em bancos um pouco mais complexos e que serão trabalhados por digitadores externos.

Finalmente, a orelha "Alerta de erro" permite que se customize a mensagem que surgirá quando da tentativa de entrada de um dado que não seja válido (veja Figs. 4.2 e 4.3).

3. Criando painéis e agilizando o processo de digitação

A seguir, a "criação de painéis" é outro importante recurso de formatação da planilha, que facilita a digitação. Como a primeira linha da planilha corresponde ao nome das variáveis, é interessante que essa primeira linha esteja sempre visível durante a navegação, o que geralmente não acontece quando se tem um banco de dados com mais de 30 elementos amostrais. O mesmo vale para a primeira coluna, a qual é sempre destinada para os códigos de identificação. Um recurso para manter linhas e colunas sempre visíveis é a criação e o congelamento de painéis.

Para essa finalidade, seleciona-se a célula que limita a linha e coluna (em geral a B2) e, no menu "Janela", escolhe-se a opção "Congelar painéis". Como resposta, o programa cria uma demarcação abaixo da primeira linha e à direita da primeira coluna (veja Fig. 4.4).

4. Utilização da ferramenta "Formulários"

Outra opção que facilita a digitação no Excel é oferecida pelo recurso de "Formulários". Clicando em "Dados" e em seguida em "Formulários", aparece uma janela como ilustrada na Fig. 4.5. Todas as células da primeira linha são identificadas como variáveis, e os dados podem ser digitados como num programa de banco de dados. Dentre as desvantagens dessa opção está o fato de que ela só permite trabalhar com no máximo 25 variáveis. Para bancos com maior número de variáveis, os dados deverão ser digitados diretamente na planilha. Além disso, ao se utilizar o modo formulário, perdem-se as informações de validação.

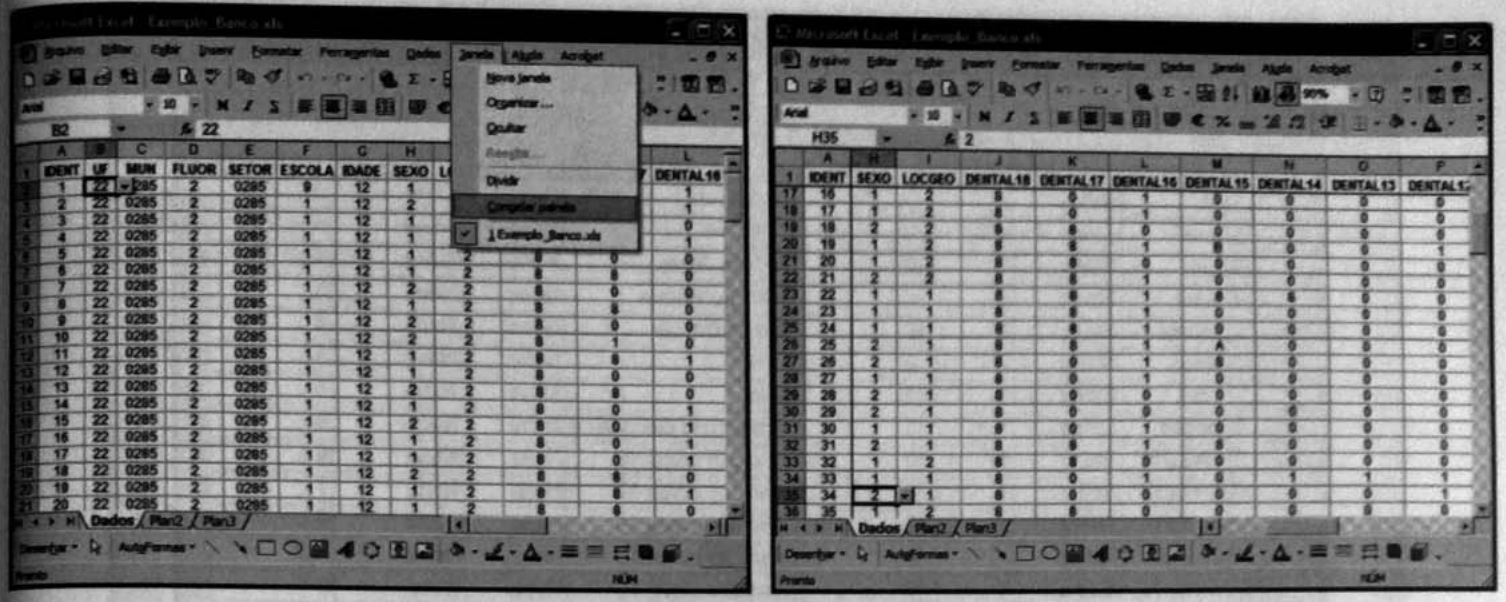


FIG. 4.4 Passos para a criação e o congelamento de painéis.

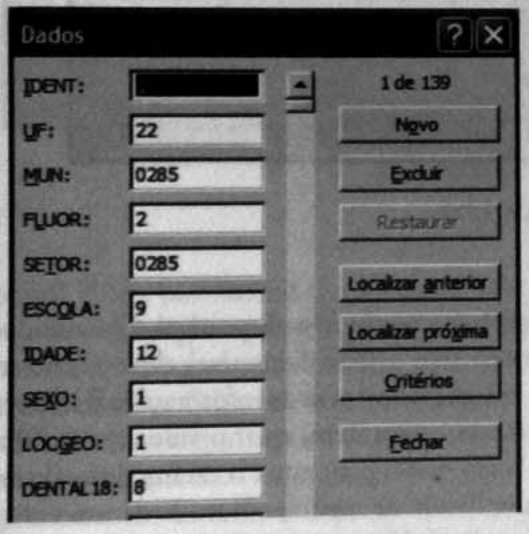


FIG. 4.5 Janela de entrada de dados da opção "Formulários".

Utilizando o SPSS®

O SPSS (Statistical Package for the Social Sciences) é um dos principais produtos da SPSS, Inc., uma empresa de software sediada em Chicago e com atividades na área de sistemas de informática desde o fim da década de 1960 (SPSS, 2004). Trata-se de um programa bastante utilizado na área acadêmica para análises estatísticas, ao lado do SAS® (Statistical Analysis System) e do Statistica®.

O SPSS tem uma interface parecida com a do Excel e permite a entrada de dados visualizando o banco de dados como um todo. Contudo, por se tratar de um programa específico para análise de dados, possui inúmeras outras potencialidades. Dentre as facilidades para operações com arquivos, suas versões mais recentes permitem ler arquivos de praticamente todos os programas mais importantes, como o próprio Excel e outras planilhas eletrônicas como Lotus, além do formato Dbase (.dbf). Sua grande desvantagem é o preço muito elevado, fator que, na maioria dos casos, restringe sua aplicação para usuários corporativos.

ETAPAS PARA A CONSTRUÇÃO DE BANCOS NO SPSS

Uma das vantagens de programas específicos de bancos de dados é facilitar a definição de variáveis. Nesse sentido, a primeira medida a se tomar é definir as variáveis. A tela de abertura do SPSS, quando se opta pela abertura de um banco de dados novo, tem duas modalidades de exibição (ou *views*): a visualização dos dados (*Data View*) e das variáveis (*Variable View*). O exemplo a seguir advém do mesmo banco ilustrado no item anterior quando discutimos o Excel.

Pode-se observar que, no *Variable View*, cada variável é definida a partir dos seguintes parâmetros (veja Fig. 4.6).

Name: Nome da variável. Conforme discutimos anteriormente, deve se limitar a oito caracteres, sem utilização de cedilhas, acentos e espaços.

Type: Tipo de variável. Existem diversos tipos disponíveis, porém os mais utilizados são o formato *String*, para variáveis categóricas, e o *Numeric*, para dados quantitativos, além de diferentes opções para o registro de datas. Trata-se de uma propriedade importante, pois irá definir a forma como o programa interpretará o dado. Uma variável do tipo *String*, por exemplo, não permite operações matemáticas nem a obtenção de medidas de tendência central e de variabilidade; para sua análise, só poderão ser obtidas freqüências.

Width: Tamanho do campo. Deve ser informado com quantos caracteres é formada cada categoria da variável. Por exemplo, se estamos trabalhando com renda e o máximo encontrado foi de 20 mil reais, então o campo deverá ter 5 algarismos. Embora colocar um tamanho maior que o necessário não atrapalhe a análise, é importante se ater ao número