Construção de banco de dados (versão 1.1)

Bancos de dados são arquivos nos quais ficam registrados os dados obtidos em uma coleta de dados de uma pesquisa. Bancos de dados bem elaborados garantem eficiência na etapa analítica da pesquisa. Os dados registrados em bancos de dados incluem informações sobre o perfil sociodemográfico, questionários, escalas, instrumentos psicológicos, respostas textuais, etc. Os bancos de dados podem apresentar variações quanto a sua organização dependendo do desenho metodológico da pesquisa. A seguir serão apresentadas instruções gerais para a construção de bancos de dados.

Leia todo o documento antes de iniciar a construção de seu banco de dados.

1) Formato, softwares e extensões

Em geral, os bancos de dados possuem o formato de uma tabela, em que as linhas representam os casos ou observações individuais, e as colunas representam as variáveis investigadas. Veja o exemplo:

Identificação	Idade	Sexo	Nível educacional	...
id1	21	Masculino	Fundamental	...
id2	32	Feminino	Médio	...
id3	18	Feminino	Superior	...
id4	46	Masculino	Superior	...
...	...	...	...	...

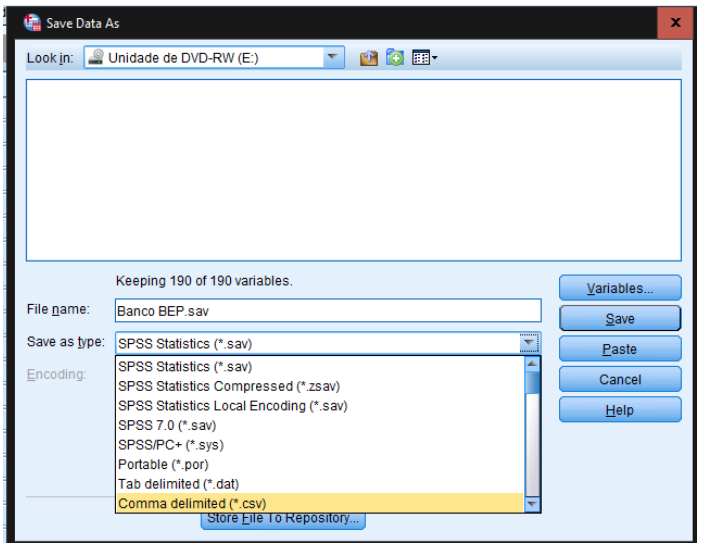
Nota. As reticências "..." significam a continuação de valores de casos ou variáveis.

As colunas representam as variáveis (Identificação, Idade, Sexo e Nível Educacional). As linhas representam cada participante da pesquisa. Eles são identificados, o que permite a conferência entre o banco de dados e os protocolos originais de coleta (no caso de coleta "caneta e papel", prontuários, instrumentos, etc.). A identificação dos casos ainda permite a recuperação da ordem original dos registros, caso a ordem seja alterada por algum motivo.

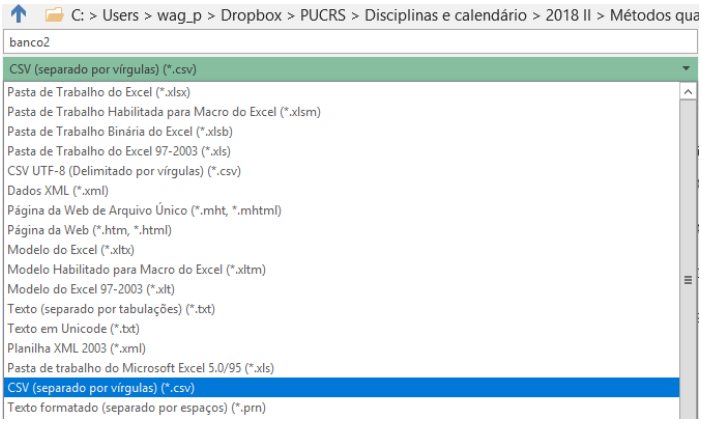
Em relação aos softwares, os mais comuns para a construção de bancos de dados são o Microsoft Excel® e o Statistical Package for the Social Sciences – SPSS®, contudo, qualquer editor de texto ou de tabelas pode ser utilizado (como as versões disponíveis no LibreOffice® ou OpenOffice®).

Quanto às extensões dos arquivos, existem muitas possibilidades. É sugerido a extensão comma-separated values (“.csv”), que organiza os dados em um formato em que todos os valores no banco de dados são separados por um caractere especial, o ponto e vírgula (“;”). O ponto e vírgula (“;”) não designa separadores de decimais e classes de algarismos, como no caso de vírgulas e pontos finais. Desta forma, você pode substituir pontos por vírgulas, e vice-versa, sem se preocupar em misturar os valores de diferentes variáveis (alguns softwares usam ponto final para separar classes de algarismos e vírgula para separar decimais, outros o inverso). Para salvar um banco de dados no formato “.csv” nos softwares mais utilizados, faça da seguinte forma:

SPSS:



Excel:



Banco de dados no formato ".csv"

```
Identificação;Idade;Sexo;Nível educacional
id1;21;Masculino;Fundamental
id2;32;Feminino;Médio
id3;18;Feminino;Superior
id4;46;Masculino;Superior
```

2) Renomear e rotular variáveis, transformar valores

Para evitar problemas de visualização e facilitar a localização das variáveis em um banco de dados, é necessário renomeá-las. Utilizando regras simples (e.g. as iniciais do nome da variável), é possível modificar os nomes das variáveis como, por exemplo, "Identificação" para "ID". Este procedimento evita sobreposições, cortes e distorções ao visualizar seu banco de dados ou criar gráficos. Ainda, é necessário pode criar rótulos e legendas para as variáveis e valores de modo a preservar uma descrição completa de ambas. A transformação de alguns valores para algarismos se faz também necessária. Em algumas análises os valores das variáveis precisam estar representados por algarismos, por exemplo, a variável "Sexo", com valores de "1" para "Feminino" e "0" para "Masculino". Observe o exemplo de um banco de dados com variáveis e valores renomeados e seus respectivos rótulos:

Banco renomeado:

ID	Idd	Sex	NEd	...
id1	21	0	1	...
id2	32	1	2	...
id3	18	1	3	...
id4	46	0	3	...
...	...	...	...	...

Rótulos e legendas:

ID	Identificação	
Idd	Idade	
Sex	Sexo	0 = Masculino, 1 = Feminino
NEd	Nível educacional	1 = Fundamental, 2 = Médio, 3 = Superior, 4 = pós-graduação, 5 = mestrado, 6 = doutorado

- Dica: Ao tomar sua decisão sobre qual número atribuir a uma categoria de resposta de uma variável, procure fazê-la de forma racional. Por exemplo, se há uma variável do tipo "possui animal de estimação", atribua "0" para "não" e "1" para "sim". Com a variável "nível educacional", atribua de forma crescente "1"

para “fundamental”, “2” para “médio”, “3” para “superior”, e assim por diante. Desta forma será mais fácil interpretar seus resultados, pois está preservado o significado dos números para representar categorias e valores. Em alguns casos, como na variável “Sexo”, esta decisão é, em geral, arbitrária.

3) Lógica de nomeação de variáveis

Por fim, outro aspecto relevante na criação de bancos de dados é a lógica de nomeação das variáveis. Imagine a situação hipotética na qual o pesquisador coletou dados de uma “Escala de Felicidade”, respondida em uma escala tipo Likert de três pontos: “0” = “discordo”, “1” = “não sei” e “2” = “concordo”. A variável contendo o escore geral da “Escala de Felicidade” pode ser nomeada como “EF”, e os itens da escala como “EF01”, “EF02”, e assim sucessivamente. Esta forma de registro de cada item de uma escala, questionário ou outro instrumento, é muito importante. O pesquisador pode estar interessado em análises específicas sobre a estrutura interna de instrumentos e outras medidas, e não apenas no seu escore geral.

- Dica: evitar caracteres especiais (ponto, vírgula, “;”, “#”, “$”, “%”, “*”, “-”, etc.) ao nomear variáveis. Alguns editores de dados simplesmente não reconhecem esses caracteres. Use, no máximo, o traço baixo “_”.

O exemplo de registro no banco de dados ficaria assim:

ID	Idd	Sex	NEd	...	EF01	EF02	...	EF	...
id1	21	0	1	...	1	1	...	9	...
id2	32	1	2	...	2	1	...	14	...
id3	18	1	3	...	0	1	...	5	...
id4	46	0	3	...	1	2	...	12	...
...	...	...	...	...	...	...	...	...	...

E os rótulos e legendas, assim:

ID	Identificação	
Idd	Idade	
Sex	Sexo	0 = Masculino, 1 = Feminino
NEd	Nível educacional	1 = Fundamental, 2 = Médio, 3 = Superior, 4 = pós-graduação, 5 = mestrado, 6 = doutorado
EF01	“conteúdo do item 1 da Escala de Felicidade”	0 = discordo, 1 = não sei, 3 = concordo
...	...	...
EF	Escala de Felicidade	Somatório da escala

- Dica: Ao nomear as variáveis em um banco de dados, procure a forma mais prática para poder localizá-las posteriormente. Por exemplo, se o seu "Questionário Sociodemográfico" possui a segunda questão sobre "Idade", é mais prático nomear essa variável no seu banco como "idade" ou "idd", ao invés de "QSD02". Uma estratégia também importante é numerar a posição da variável no banco. Por exemplo, se o primeiro item de uma "Escala de Depressão" fica localizado na vigésima coluna do banco de dados, é possível nomeá-la como "ED01_20". Com este procedimento é possível indicar tanto a qual questionário ou instrumento a variável/item pertence, quanto indicar que ela é a vigésima coluna do banco de dados. Alguns pacotes estatísticos, como o R, utilizam o número da coluna para selecionar variáveis em sua sintaxe. Assim é possível poupar o trabalho de contar colunas no momento de preparar a análise dos dados.

Edição 1: dados faltantes ou *missing values*

Os dados faltantes, ou *missing values*, podem ser tratados de duas formas. Na primeira forma, mais utilizada, o pesquisador define um valor muito diferente dos possíveis valores de suas variáveis para designar um dado faltante. Pode-se utilizar 99 ou 999. Muito importante: registrar isto em sua planilha de rótulos e legendas. Na segunda forma, basta deixar a célula da planilha em branco ou atribuir o valor "NA", que no software R é lido como *not available*.

Feedback e sugestões: wagner.machado@pucrs.br