Sugestões para construção de planilhas de dados

Julio M. Singer e Pedro A. Morettin

Departamento de Estatística, IME/USP

Planilhas (usualmente eletrônicas) são matrizes em que se armazenam dados com o objetivo de permitir sua análise estatística. Em geral, cada linha da matriz de dados corresponde a uma unidade de investigação (*e.g.* unidade amostral) e cada coluna, a uma variável. A primeira etapa para a construção de uma planilha de dados consiste na elaboração de um dicionário com a especificação das variáveis, que envolve

i) sua definição operacional;

ii) a atribuição de rótulos (mnemônicos e com letras minúsculas para facilitar a digitação);

iii) a especificação das unidades de medida ou definição de categorias;

iv) a atribuição de um código para valores omissos (*missing*);

v) a escolha entre ponto ou vírgula para separação de casas decimais[1];

vi) a especificação do número de casas decimais (correspondente à precisão do instrumento de medida).

Algumas recomendações para a construção da planilha de dados são

i) não utilizar limitadores de celas (*borders*) ou cores;

ii) reservar a primeira linha para os rótulos das variáveis;

iii) não esquecer uma coluna para a variável indicadora das unidades de investigação.

iv) atribuir um código para valores omissos (*missing*);

v) escolher ponto ou vírgula para separação de casas decimais;

vi) especificar o número de casas decimais (correspondente à precisão do instrumento de medida).

Exemplo 1 Os dados da Tabela 2 foram extraídos de um estudo realizado no Instituto de Ciências Biomédicas da Universidade de São Paulo com o objetivo de avaliar a associação entre a infecção de gestantes por malária e a ocorrência de microcefalia nos respectivos bebês. O dicionário das variáveis observadas está indicado na Tabela 1.

A disposição dos dados do Exemplo 1 no formato de uma planilha está representada na Tabela 2.

[1]Embora a norma brasileira ABNT indique a vírgula para separação de casas decimais, a maioria dos pacotes computacionais utiliza o ponto com essa função; por essa razão é preciso tomar cuidado com esse detalhe na construção de planilhas a serem analisadas computacionalmente. Em geral adotaremos a norma brasileira no texto.

Tabela 1: Dicionário para as variáveis referentes ao estudo

Rótulos	**Variável**	**Unidade de medida**
idade	Idade da mãe	anos
nummal	Quantidade de malárias durante a gestação	número inteiro
parasita	Espécie do parasita da malária	0: não infectada
		1: P. vivax
		2: P. falciparum
		3: malária mista
		4: indeterminado
numgest	Paridade (quantidade de gestações)	Número inteiro
idgest	Idade gestacional no parto	semanas
sexorn	Sexo do recém-nascido	1: masculino
		2: feminino
pesorn	Peso do recém-nascido	g
estrn	Estatura do recém-nascido	cm
pcefal	Perímetro cefálico do recém-nascido	cm
Obs:	Observações omissas são representadas por um ponto	

Exemplo 2 Na Tabela 3 apresentamos dados provenientes de um estudo em que o objetivo é avaliar a variação do peso (kg) de bezerros submetidos a uma determinada dieta entre 12 e 26 semanas após o nascimento.

Dados com essa natureza são chamados de dados longitudinais por terem a mesma característica (peso, no exemplo) medida ao longo de uma certa dimensão (tempo, no exemplo). De acordo com nossa especificação, há nove variáveis na planilha representada na Tabela 3, nomeadamente, Animal, Peso na 12a semana, Peso na 14a semana etc. Para efeito computacional, no entanto, esse tipo de dados deve ser disposto numa planilha com formato diferente (às vezes chamado de formato longo) como indicado na Tabela 4.

Nesse formato apropriado para dados longitudinais (ou mais geralmente, para medidas repetidas), há apenas três variáveis, a saber, Animal, Semana e Peso. Note que a mesma unidade amostral (animal) é repetida na primeira coluna para caracterizar a natureza longitudinal dos dados. Esse formato também é comumente utilizado para armazenar dados de séries temporais.

Tabela 2: Planilha com dados referentes ao Exemplo 1

ident	idade	nummal	parasita	numgest	idgest	sexorn	pesorn	estrn	pcefal
1	25	0	0	3	38	2	3665	46	36
2	30	0	0	9	37	1	2880	44	33
3	40	0	0	1	41	1	2960	52	35
4	26	0	0	2	40	1	2740	47	34
5	.	0	0	1	38	1	2975	50	33
6	18	0	0	.	38	2	2770	48	33
7	20	0	0	1	41	1	2755	48	34
8	15	0	0	1	39	1	2860	49	32
9	.	0	0	.	42	2	3000	50	35
10	18	0	0	1	40	1	3515	51	34
11	17	0	0	2	40	1	3645	54	35
12	18	1	1	3	40	2	2665	48	35
13	30	0	0	6	40	2	2995	49	33
14	19	0	0	1	40	1	2972	46	34
15	32	0	0	5	41	2	3045	50	35
16	32	0	0	8	38	2	3150	44	35
17	18	0	0	2	40	1	2650	48	33.5
18	18	0	0	1	41	1	3200	50	37
19	19	0	0	1	39	1	3140	48	32
20	18	0	0	2	40	1	3150	47	35
21	27	0	0	3	40	1	4185	52	35.5
22	26	0	0	3	40	2	4070	52	35
23	.	0	0	.	40	1	3950	50	37
24	19	0	0	1	40	1	3245	51	33
25	23	0	0	.	41	1	3010	49	35
26	.	0	0	.	40	2	3260	50	33
27	20	1	1	2	40	2	3450	49	33
28	19	0	0	3	40	2	2765	48	32
29	22	0	0	4	40	1	4190	50	34
30	32	0	0	4	42	2	4035	51	34
31	33	0	0	5	39	2	3620	51	33
32	30	3	3	5	38	1	3230	48	34
33	36	0	0	7	39	2	3185	50	38
34	.	0	0	.	39	2	2950	47	33

Tabela 3: Peso de bezerros (kg)

	Semanas após nascimento							
animal	12	14	16	18	20	22	24	26
1	54.1	65.4	75.1	87.9	98.0	108.7	124.2	131.3
2	91.7	104.0	119.2	133.1	145.4	156.5	167.2	176.8
3	64.2	81.0	91.5	106.9	117.1	127.7	144.2	154.9
4	70.3	80.0	90.0	102.6	101.2	120.4	130.9	137.1
5	68.3	77.2	84.2	96.2	104.1	114.0	123.0	132.0
6	43.9	48.1	58.3	68.6	78.5	86.8	99.9	106.2
7	87.4	95.4	110.5	122.5	127.0	136.3	144.8	151.5
8	74.5	86.8	94.4	103.6	110.7	120.0	126.7	132.2
9	50.5	55.0	59.1	68.9	78.2	75.1	79.0	77.0
10	91.0	95.5	109.8	124.9	135.9	148.0	154.5	167.6
11	83.3	89.7	99.7	110.0	120.8	135.1	141.5	157.0
12	76.3	80.8	94.2	102.6	111.0	115.6	121.4	134.5
13	55.9	61.1	67.7	80.9	93.0	100.1	103.2	108.0
14	76.1	81.1	84.6	89.8	97.4	111.0	120.2	134.2
15	56.6	63.7	70.1	74.4	85.1	90.2	96.1	103.6

Tabela 4: Planilha computacionalmente adequada para os dados do Exemplo 2

animal	semana	peso
1	12	54.1
1	14	65.4
1	16	75.1
1	18	87.9
1	20	98.0
1	22	108.7
1	24	124.2
1	26	131.3
2	12	91.7
2	14	104.0
⋮	⋮	⋮
2	26	176.8
⋮	⋮	⋮
15	12	56.6
⋮	⋮	⋮
15	26	103.6