

# *Organização dos dados*

*Rodrigo Citton P. dos Reis, Dep. de Estatística - UFRGS*

*Agosto de 2020*

## *Introdução*

Agora que já discutimos alguns **conceitos básicos** de estatística e as etapas gerais de um **levantamento estatístico**, vamos apresentar como é feito o **registro** e a **organização de dados** referentes a uma certa coleta de dados. Começaremos com a **planilha** para o registro dos dados e a **tabela de dados brutos** resultante. Logo em seguida, discutiremos como fazer a **apuração dos dados**.

## *Coleta de dados*

- **Lembrando:** a **estatística** é a ciência que tem por objetivo orientar a *coleta*, o *resumo*, a *apresentação*, a *análise* e a *interpretação* de dados.

Para *coletar dados*, o pesquisador necessitará armazenar os dados coletados em algum lugar. Assim, se faz necessário organizar uma **planilha**. Com o advento da computação, grande parte dos profissionais da área de estatística registram dados em uma *planilha eletrônica*<sup>1</sup>. No entanto, os dados também podem ser registrados em meio físico como, por exemplo, fichas, cadernos ou cadernetas, ou seja, a chamada *planilha física*.

As planilhas eletrônicas podem ser construídas a partir de planilhas físicas ou serem alimentadas por algum **instrumento de coleta** em meio eletrônico (formulário ou questionário)<sup>2</sup>. Vamos apresentar como se desenha uma planilha física para registro dos dados. Se você tiver possibilidade, pode experimentar como organizar os dados em uma planilha eletrônica.

- **Planilha** é o documento que armazena os dados coletados, distribuindo-os em linhas e colunas<sup>3</sup>. Em planilhas eletrônicas, geralmente, as linhas são numeradas e as colunas são indicadas por letras maiúsculas.

**Exemplo:** considere o exemplo adaptado de [Morettin and Bussab, 2017]. Um pesquisador está interessado em fazer um levantamento sobre alguns aspectos socioeconômicos dos empregados da seção de orçamentos da Companhia MB, um grupo de 15 pessoas. Temos a seguinte planilha (Fig. 2) para registrar os dados do grupo.

- **Dados brutos** são os dados na forma em que foram coletados, sem qualquer tipo de tratamento.

<sup>1</sup> Softwares como *Calc* (OpenOffice), *Microsoft Excel* (Office) e *Google Sheets* (Google) são exemplos de *softwares* que trabalham com planilhas eletrônicas.

<sup>2</sup> O *Google Forms*, por exemplo, cria e alimenta uma planilha eletrônica a partir do formulário de coleta.

<sup>3</sup> Ou seja, planilhas são “matrizes de dados”.

	A	B	C	D
1				
2				D2
3				
4				
5				

Figure 1: Célula D2, no cruzamento da coluna D com a linha 2.

Número	Estado Civil	Grau de instrução	Número de filhos	Salário (x sal. mín.)	Idade	Região de procedência
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						

Figure 2: Planilha física para o registro dos dados do grupo de 15 empregados da seção de orçamentos da Companhia MB.

Após a coleta de dados, o pesquisador tem em sua planilha o registro dos dados brutos (Fig. 3).

- O que podemos falar sobre as variáveis coletadas?
- Qual a informação podemos apresentar sobre os dados coletados?

Para responder tais perguntadas, precisaremos **resumir** os dados de alguma forma. Na próxima seção discutiremos a etapa de **apuração dos dados**.

- **Exercício:** construa a planilha para o registro do levantamento dos dados planejado nas aulas anteriores.

### Apuração dos dados

- **Apuração** é o processo de retirar os dados da planilha e organizá-los, para apresentação.

No exemplo apresentado anteriormente, foram coletadas as seguintes variáveis: estado civil, grau de instrução, número de filhos, salário, idade e região de procedência. Note que estas são variáveis de diferentes tipos<sup>4</sup>.

<sup>4</sup> **Exercício:** classifique cada uma destas variáveis em *qualitativa nominal*, *qualitativa ordinal*, *quantitativa discreta* e *quantitativa contínua*.

Número	Estado Civil	Grau de instrução	Número de filhos	Salário (x sal. mín.)	Idade	Região de procedência
1	solteiro	EF	0	4,00	26	interior
2	casado	EF	1	4,56	32	capital
3	casado	EF	2	5,25	36	capital
4	solteiro	EM	0	5,73	20	litoral
5	solteiro	EF	1	6,26	40	litoral
6	casado	EF	0	6,66	28	interior
7	solteiro	EF	2	6,86	41	interior
8	solteiro	EF	1	7,39	43	capital
9	casado	EM	1	7,59	34	capital
10	solteiro	EM	0	7,44	23	litoral
11	casado	S	2	8,12	33	interior
12	solteiro	EF	0	8,46	27	capital
13	solteiro	EM	2	8,74	37	litoral
14	casado	EF	3	8,95	44	litoral
15	casado	EM	0	9,13	30	interior

Figure 3: Planilha com o registro dos dados brutos do grupo de 15 empregados da seção de orçamentos da Companhia MB. EF, EM e S representam Ensino Fundamental, Ensino Médio e Superior, respectivamente.

### Apuração de dados nominais

Se quisermos saber quantos solteiros e quantos casados trabalham na seção de orçamentos da Companhia MB devemos escrever os valores possíveis da variável *estado civil*<sup>5</sup>.

solteiro  
casado

Logo após, precisamos inspecionar cada registro da tabela de dados brutos e marcar um traço ao lado de *solteiro*, para cada indivíduo solteiro inspecionado, e um traço ao lado de *casado* para cada indivíduo casado inspecionado. A cada quatro traços, corta-se com um traço, e este conjunto representa uma contagem de cinco indivíduos <sup>6</sup>.

solteiro    |||| |||  
casado    |||| ||

Desta forma, verificamos que na seção de orçamentos da Companhia MB trabalham oito solteiros e sete casados. Duas outras formas alternativas de se fazer a apuração dos dados são apresentadas a seguir<sup>7</sup>.

<sup>5</sup> **Pergunta:** a ordem de escrita dos valores possíveis da variável *estado civil* importa? Por que?

<sup>6</sup> No inglês, *tally marks* (marcas de registro).

<sup>7</sup> **Comentário:** é fácil apurar uma pequena massa de dados, como no caso do exemplo. Já uma grande massa de dados tornará a tarefa difícil e entediante. Além disso, com um grande volume de dados, a *probabilidade* de incorrermos em erros aumenta! Necessitaremos do auxílio de pacotes estatísticos!

solteiro    
casado

solteiro    
casado

### Apuração de dados ordinais

Para apurar dados de grau de instrução (variável qualitativa ordinal), o procedimento é similar ao adotado para apurar dados nominais. A diferença é que, para dados ordinais, **impõe-se uma ordem**. Contudo, a apuração se faz por contagem.

Ensino Fundamental		= 5
Ensino Médio		= 4
Superior		= 1

### Apuração de dados discretos

Para apurar o número de filhos (variável quantitativa discreta), também devemos fazer uma contagem. Escrevemos os resultados respeitando a ordem numérica.

0		= 6
1		= 4
2		= 4
3		= 1

### Apuração de dados contínuos

Em geral, os dados contínuos são apresentados na forma como foram coletados, porque assumem valores diferentes, mesmo em amostras

pequenas. É o caso da variável idade no exemplo considerado: os empregados da seção de orçamentos da Companhia MB tinham idades diferentes. No entanto, é possível organizar as idades por faixas, como veremos nas aulas seguintes.

### *Exercícios*

Faça uma pequena coleta de dados incluindo pelo menos uma variável de cada tipo (*qualitativa nominal*, *qualitativa ordinal*, *quantitativa discreta* e *quantitativa contínua*).

1. Organize uma planilha (física ou eletrônica) para o registro dos dados coletados.
2. Faça a coleta e preencha a planilha para obter os dados brutos.
3. Faça a apuração dos dados e comente brevemente sobre os resultados encontrados.

### *ComplementaR*

Esta seção é complementar. São apresentadas algumas poucas funções em R relacionadas a discussão da aula. Para tal, vamos utilizar o exemplo original de [Morettin and Bussab, 2017] sobre os dados dos empregados da seção de orçamentos da Companhia MB. A planilha eletrônica correspondente encontra-se no arquivo `companhia_mb.xlsx`. Vamos começar carregando os dados para o R. Existem várias formas de se carregar **arquivos de dados** em diferentes no R. Como arquivo de interesse encontra-se no formato do Excel (xlsx), vamos utilizar a função `read_excel` do pacote `readxl`<sup>8</sup>.

```
# install.packages("readxl")
library(readxl)

dados <- read_excel(path = "companhia_mb.xlsx")

class(dados) # classe do objeto dados

## [1] "tbl_df"     "tbl"        "data.frame"

dim(dados) # dimensão do objeto dados

## [1] 36  7
```

Note que o objeto `dados` é uma tabela de dados bruto.

```
head(dados) # apresenta as primeiras linhas do objeto dados
```

<sup>8</sup> Caso você não tenha o pacote, instale-o:  
`install.packages("readxl")`.

```

## # A tibble: 6 x 7
##   N `Estado Civil` `Grau de Instru~ `N de Filhos` `Salario (x Sal~ Idade
##   <dbl> <chr>          <chr>                  <dbl>          <dbl> <dbl>
## 1     1 solteiro      ensino fundamen~       NA            4    26
## 2     2 casado        ensino fundamen~       1            4.56   32
## 3     3 casado        ensino fundamen~       2            5.25   36
## 4     4 solteiro      ensino médio         NA            5.73   20
## 5     5 solteiro      ensino fundamen~       NA            6.26   40
## 6     6 casado        ensino fundamen~       0            6.66   28
## # ... with 1 more variable: `Região de Procedência` <chr>

```

A função `table` retorna contagens dos valores de cada variável, e portanto, podemos utilizar esta função para a apuração dos dados.

```



```

## References

Pedro A. Morettin and Wilton de O. Bussab. *Estatística básica*.  
 Saraiva, São Paulo, 9 edition, July 2017. ISBN 978-85-472-2022-8.