Universidade Estadual de Campinas – UNICAMP

Hospital da Mulher Prof. Dr. José Aristodemo Pinotti – CAISM

Núcleo de Estatística – CAISM / UNICAMP

MANUAL para CONSTRUÇÃO de BANCO de DADOS

Versão 1

Elaborado pelos Estatísticos

Graciella Dalla Torre

Marcelo Tavares de Lima

CONSTRUÇÃO de BANCO de DADOS

1. Introdução

Este documento apresenta orientações gerais para elaboração de **banco de dados** em planilhas eletrônicas, mais especificamente, em planilhas Microsoft Excel®. Sua elaboração se motiva na necessidade da rápida obtenção de resultados estatísticos para subsidiar pesquisas clínicas realizadas no Departamento de Tocoginecologia do Hospital da Mulher Prof. Dr. José Aristodemo Pinotti – CAISM.

O intuito deste documento é mostrar, de forma simples, como utilizar as ferramentas do Excel **para planejar a organização do banco de dados antes mesmo da coleta.**

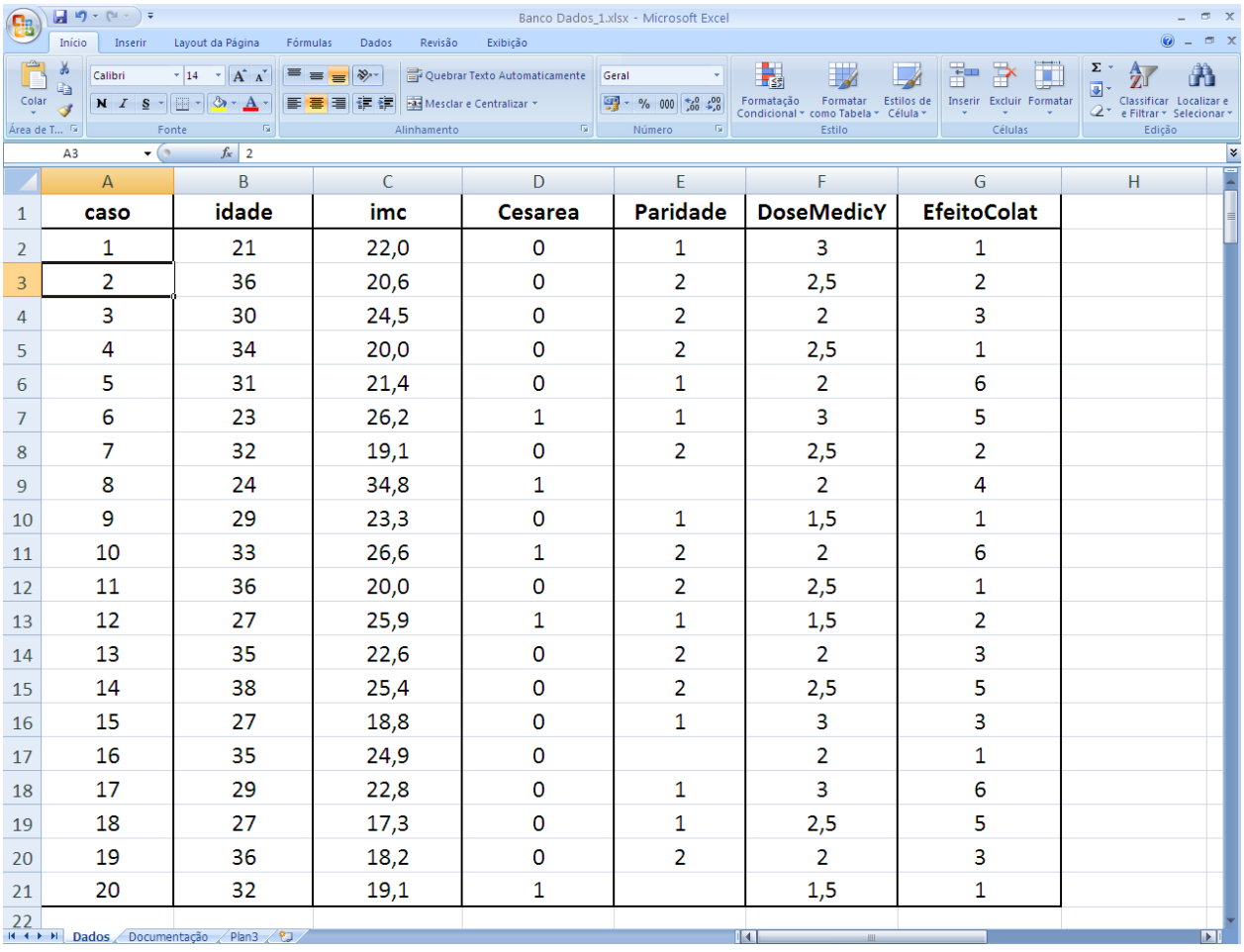
Geralmente, ao final do processo de uma pesquisa clínica, a maior parte das informações sobre os sujeitos e as variáveis estará em um **banco de dados** computadorizado que servirá para armazenar, atualizar e monitorar os dados, bem como para formatá-los para análises estatísticas.

2. Tabela de dados

Todas as bases de dados computadorizadas são compostas por uma ou mais tabelas onde as **linhas** correspondem a **registros** individuais (que podem representar sujeitos, eventos, ou transações), e as **colunas** correspondem a **campos** (“atributos” dos registros). Por exemplo, os bancos de dados mais simples consistem em uma tabela única onde cada linha corresponde a um determinado sujeito do estudo e cada coluna corresponde a um atributo específico do sujeito, como nome, data de nascimento, sexo e o valor de uma variável preditora ou de desfecho. Em geral, a primeira coluna corresponde a um **número de identificação único do sujeito** denominado **chave principal** da tabela.

A **Figura 1** mostra uma tabela de dados de um estudo clínico. Cada linha da tabela corresponde a um determinado sujeito do estudo, e cada coluna corresponde a um atributo desse sujeito.

Figura 1 – Banco de dados no Excel

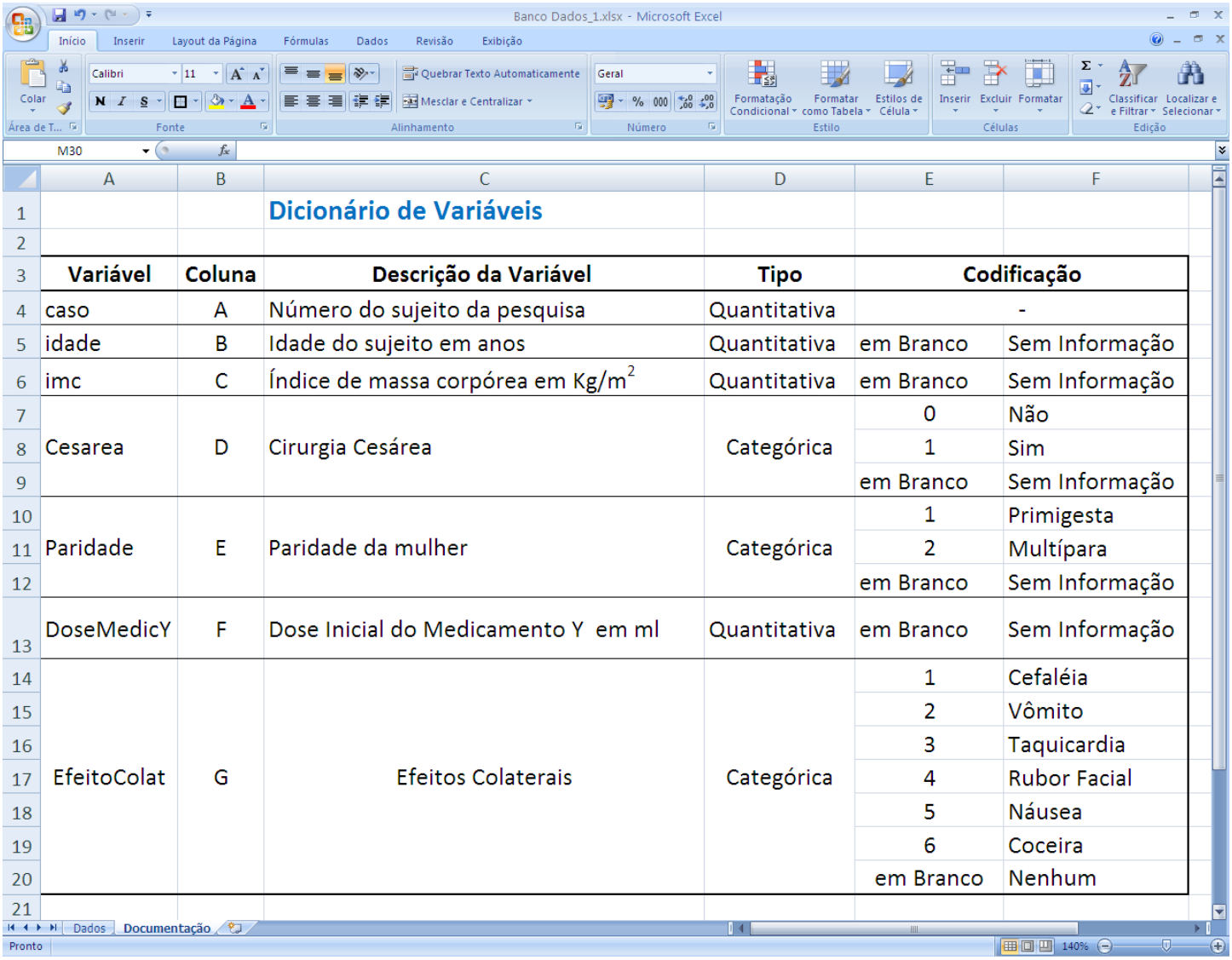


	A	B	C	D	E	F	G	H
1	caso	idade	imc	Cesarea	Paridade	DoseMedicY	EfeitoColat	
2	1	21	22,0	0	1	3	1	
3	2	36	20,6	0	2	2,5	2	
4	3	30	24,5	0	2	2	3	
5	4	34	20,0	0	2	2,5	1	
6	5	31	21,4	0	1	2	6	
7	6	23	26,2	1	1	3	5	
8	7	32	19,1	0	2	2,5	2	
9	8	24	34,8	1		2	4	
10	9	29	23,3	0	1	1,5	1	
11	10	33	26,6	1	2	2	6	
12	11	36	20,0	0	2	2,5	1	
13	12	27	25,9	1	1	1,5	2	
14	13	35	22,6	0	2	2	3	
15	14	38	25,4	0	2	2,5	5	
16	15	27	18,8	0	1	3	3	
17	16	35	24,9	0		2	1	
18	17	29	22,8	0	1	3	6	
19	18	27	17,3	0	1	2,5	5	
20	19	36	18,2	0	2	2	3	
21	20	32	19,1	1		1,5	1	
22								

No banco de dados para análise estatística, todo o conteúdo, com exceção do nome das variáveis, deve ser numérico. Se a variável for uma medida contínua, o conteúdo pode conter vírgula com casas decimais (quantas forem necessárias). **Se a variável for categórica, deve-se numerar com códigos suas categorias e deixar "em Branco" a categoria de ausência de informação.** Por exemplo, para a variável local de parto, pode-se usar "1" para "Caism" e "2" para "outros locais"

O nome das variáveis não podem: conter acentuação (por exemplo, "cesárea"),ser formado por mais de uma palavra (por exemplo "local do parto"), nem começar com número. Por isso, uma documentação conhecida como "Dicionário" deve ser criada em uma planilha à parte no mesmo arquivo Excel, identificando completamente as variáveis do banco de dados. A **Figura 2** mostra um exemplo de documentação de banco de dados.

Figura 2 – Documentação de um banco de dados



Variável	Coluna	Descrição da Variável	Tipo	Codificação	
caso	A	Número do sujeito da pesquisa	Quantitativa	-	
idade	B	Idade do sujeito em anos	Quantitativa	em Branco	Sem Informação
imc	C	Índice de massa corpórea em Kg/m^2	Quantitativa	em Branco	Sem Informação
Cesarea	D	Cirurgia Cesárea	Categórica	0	Não
				1	Sim
				em Branco	Sem Informação
Paridade	E	Paridade da mulher	Categórica	1	Primigesta
				2	Multípara
				em Branco	Sem Informação
DoseMedicY	F	Dose Inicial do Medicamento Y em ml	Quantitativa	em Branco	Sem Informação
EfeitoColat	G	Efeitos Colaterais	Categórica	1	Cefaléia
				2	Vômito
				3	Taquicardia
				4	Rubor Facial
				5	Náusea
				6	Coceira
				em Branco	Nenhum

3. EXEMPLO - Bancos de dados Errado e Correto

A **Figura 3** mostra uma tabela de dados com vários tipos de Erros na inserção dos dados. Cada linha da tabela corresponde a um determinado sujeito do estudo, e cada coluna corresponde a uma característica desse sujeito.

Figura 3 – Tabela de Dados Errada.

TABELA de DADOS - ERRADA

caso	Idade	IMC	Cesárea	Paridade	Dose do Medic Y	Efeito Colateral	Quais Efeitos Colaterais
1	21 anos	22,0 kg/m^2	Não	Primigesta	3 ml	Sim	Cefaléia,Náusea
2	36 anos	28,6 kg/m^2	Não	Multípara	2,5 ml	Sim	Vômito,Taquicardia
3	30 anos	24,5 kg/m^2	Não	Multípara	2 ml	Não	Nenhum
4	34 anos	20,0 kg/m^2	Não	Multípara	2,5 ml	Sim	Cefaléia,Náusea,Taquicardia
5	31 anos	21,4 kg/m^2	Não	Primigesta	2 ml	Sim	Rubor Facial,Coceira,Náusea
6	23 anos	26,2 kg/m^2	Sim	Primigesta	3 ml	Sim	Cefaléia,Náusea
7	32 anos	19,1 kg/m^2	Não	Multípara	2,5 ml	Sim	Vômito,Taquicardia
8	24 anos	34,8 kg/m^2	Sim	Sem Informação	2 ml	Não	Nenhum
9	29 anos	23,3 kg/m^2	Não	Primigesta	1,5 ml	Sim	Cefaléia,Náusea,Taquicardia
10	33 anos	26,6 kg/m^2	Sim	Multípara	2 ml	Sim	Rubor Facial,Coceira,Náusea
11	36 anos	20,0 kg/m^2	Não	Multípara	2,5 ml	Não	Nenhum
12	27 anos	25,9 kg/m^2	Sim	Primigesta	1,5 ml	Sim	Vômito,Taquicardia
13	35 anos	22,6 kg/m^2	Não	Multípara	2 ml	Sim	Taquicardia,Rubor Facial,Vômito
14	38 anos	25,4 kg/m^2	Não	Multípara	2,5 ml	Sim	Cefaléia,Náusea,Taquicardia
15	27 anos	18,8 kg/m^2	Não	Primigesta	3 ml	Sim	Taquicardia,Rubor Facial,Vômito
16	35 anos	24,9 kg/m^2	Não	Sem Informação	2 ml	Sim	Cefaléia,Náusea,Taquicardia
17	29 anos	22,8 kg/m^2	Não	Primigesta	3 ml	Não	Nenhum
18	27 anos	17,3 kg/m^2	Não	Primigesta	2,5 ml	Sim	Cefaléia,Náusea
19	36 anos	18,2 kg/m^2	Não	Multípara	2 ml	Sim	Taquicardia,Rubor Facial,Vômito
20	32 anos	19,1 kg/m^2	Sim	Sem Informação	1,5 ml	Sim	Cefaléia,Náusea,Taquicardia

A **Figura 4** mostra a tabela da **Figura 3** corrigida, e nas próximas páginas será ilustrada a correção de cada variável (coluna) da Tabela.

Figura 4 – Tabela de Dados Correta.

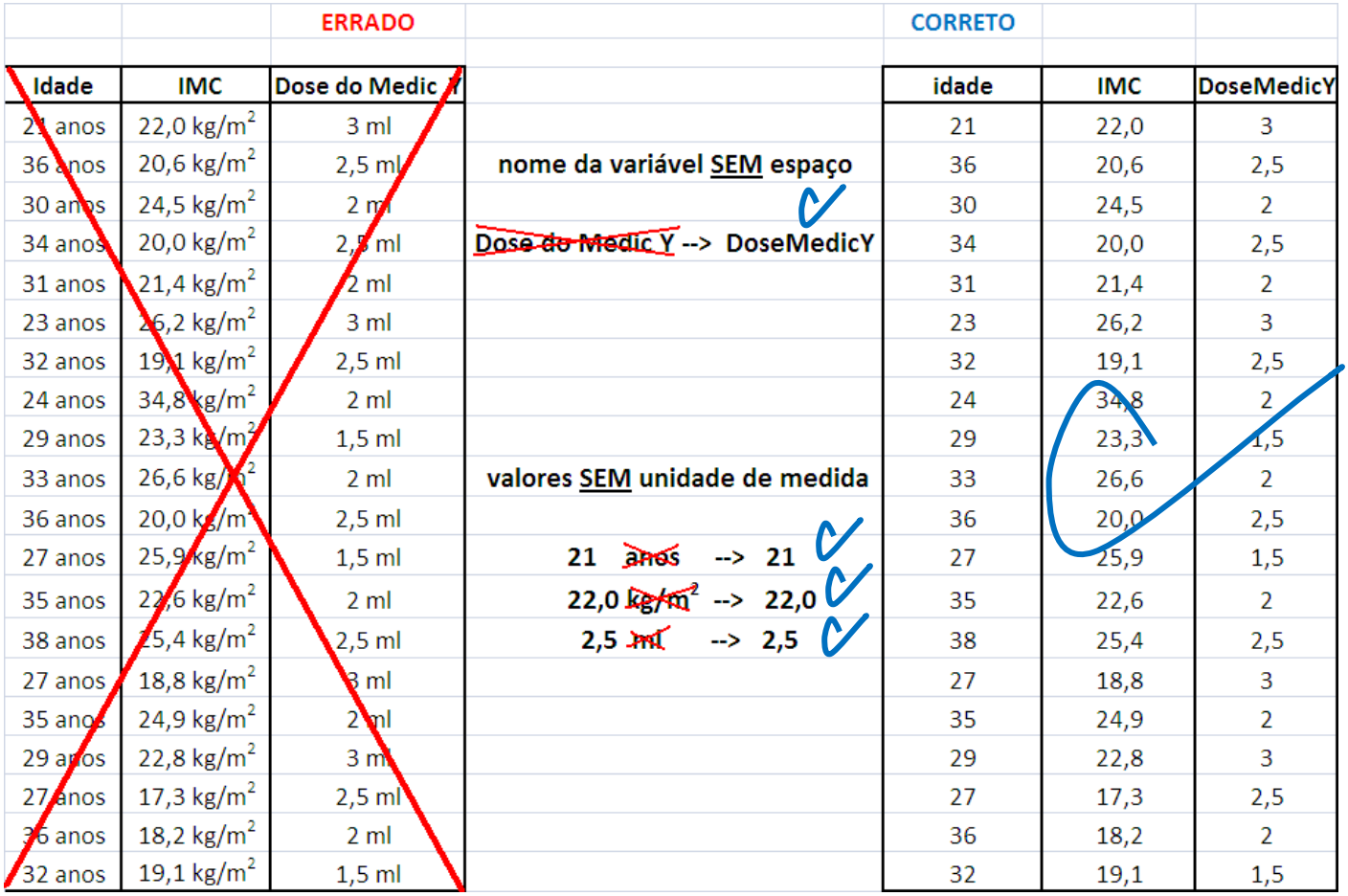
TABELA de DADOS - CORRETA

paciente	idade	IMC	Cesarea	Paridade	DoseMedicY	EfeitoColateral	Cefaleia	Nausea	Vomito	Taquicardia	RuborFacial	Coceira
1	21	22,0	0	1	3	1	1	1	0	0	0	0
2	36	20,6	0	2	2,5	1	0	0	1	1	0	0
3	30	24,5	0	2	2	0	0	0	0	0	0	0
4	34	20,0	0	2	2,5	1	1	1	0	1	0	0
5	31	21,4	0	1	2	1	0	1	0	0	1	1
6	23	26,2	1	1	3	1	1	1	0	0	0	0
7	32	19,1	0	2	2,5	1	0	0	1	1	0	0
8	24	34,8	1		2	0	0	0	0	0	0	0
9	29	23,3	0	1	1,5	1	1	1	0	1	0	0
10	33	26,6	1	2	2	1	0	1	0	0	1	1
11	36	20,0	0	2	2,5	0	0	0	0	0	0	0
12	27	25,9	1	1	1,5	1	0	0	1	1	0	0
13	35	22,6	0	2	2	1	0	0	1	1	1	0
14	38	25,4	0	2	2,5	1	1	1	0	1	0	0
15	27	18,8	0	1	3	1	0	0	1	1	1	0
16	35	24,9	0		2	1	1	1	0	1	0	0
17	29	22,8	0	1	3	0	0	0	0	0	0	0
18	27	17,3	0	1	2,5	1	1	1	0	0	0	0
19	36	18,2	0	2	2	1	0	0	1	1	1	0
20	32	19,1	1		1,5	1	1	1	0	1	0	0

4. Corrigindo as Variáveis do Banco de Dados

As Figuras que seguem abaixo são exemplos mais comuns de erros na inserção dos dados. Logo ao lado de cada erro, é ilustrada a forma correta de inserir a resposta de cada tipo de variável.

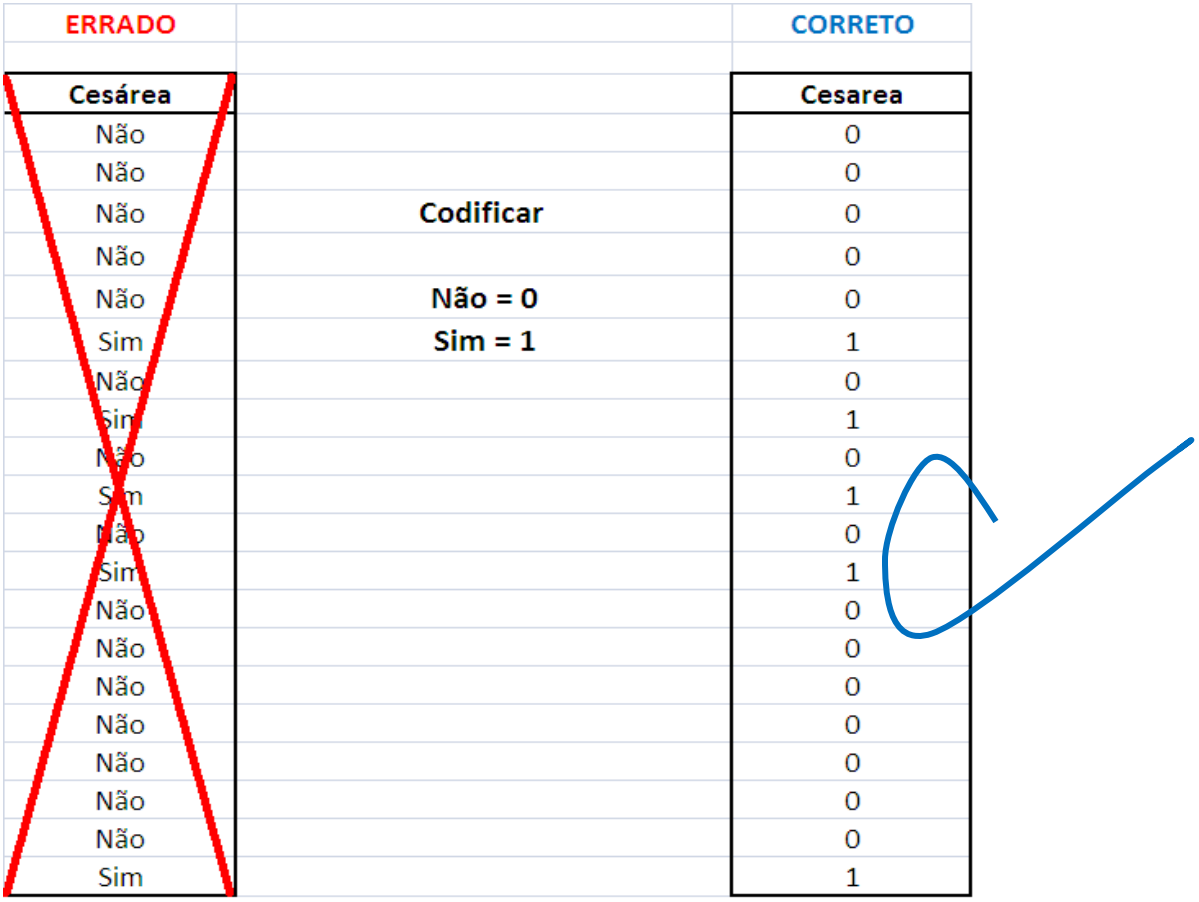
Figura 5 – Exemplo de Variáveis Quantitativas inseridas com unidade de medida.



Neste exemplo da **Figura 5** ocorre o erro de se nomear uma variável com mais de uma palavra. Portanto, o nome da variável “Dose do Medic Y” foi corrigido para “DoseMedicY”.Percebam que o nome é uma palavra única (sem espaços entre palavras) e abreviado.

Outro erro comum é inserir os valores das variáveis Quantitativas com a unidade de medida (anos, kg/m^2, ml), a correção foi realizada retirando todas as unidades de medidas, deixando apenas os números que indicam o valor da variável coletada.

Figura 6 – Exemplo de Variável Categórica dicotômica com resposta (Sim / Não)

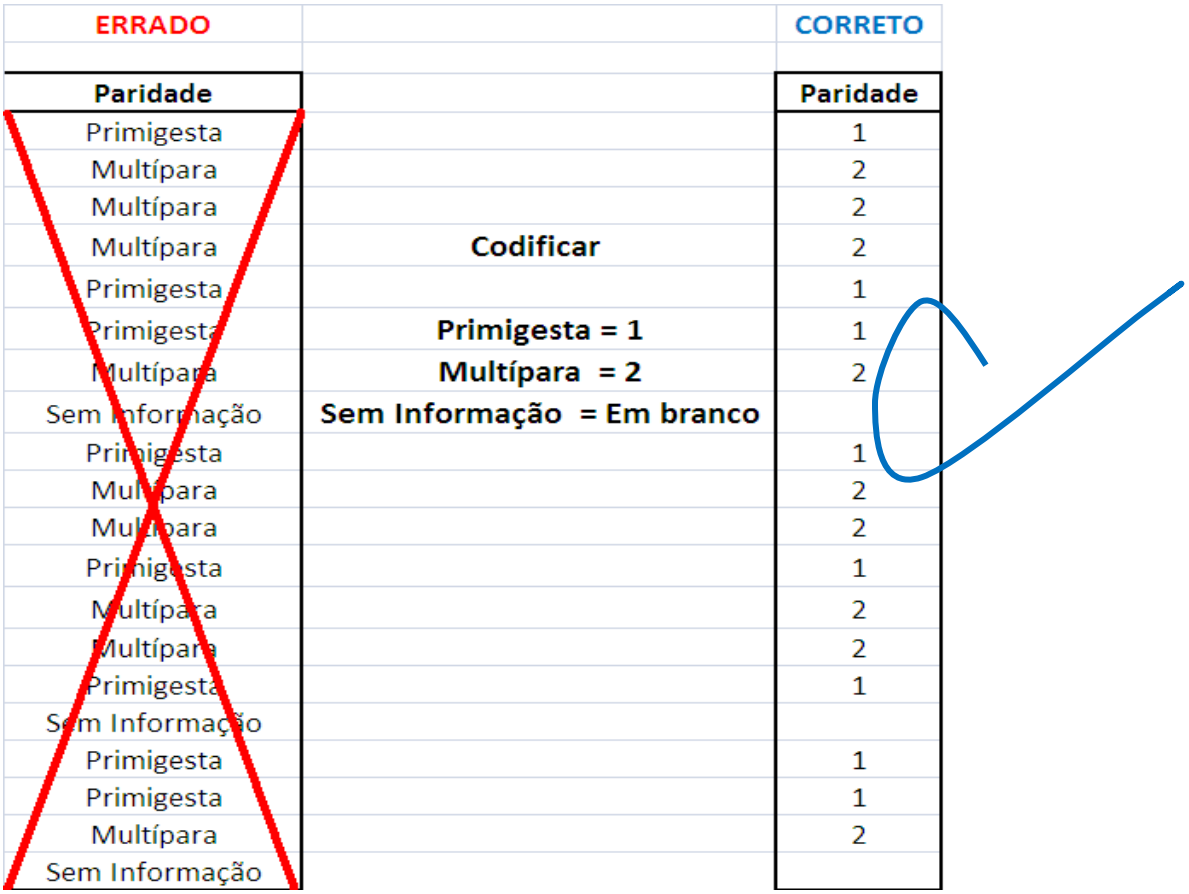


ERRADO		CORRETO
Cesárea		**Cesarea**
Não		0
Não		0
Não	**Codificar**	0
Não		0
Não	**Não = 0**	0
Sim	**Sim = 1**	1
Não		0
Sim		1
Não		0
Sim		1
Não		0
Sim		1
Não		0
Não		0
Não		0
Não		0
Não		0
Não		0
Não		0
Sim		1

Se a variável é do tipo categórica dicotômica (duas possibilidades de respostas), deve-se numerar com códigos suas categorias.

Geralmente quando se estuda uma característica Presente ou Ausente no indivíduo. Deve-se codificar com “0” (zero) quando a resposta é “Não” (característica ausente) e “1” (um) quando a resposta é “Sim” (característica Presente).

Figura 7 – Exemplo de Variável Categórica com valores faltantes



ERRADO		CORRETO
Paridade		**Paridade**
Primigesta		1
Multípara		2
Multípara		2
Multípara	**Codificar**	2
Primigesta		1
Primigesta	**Primigesta = 1**	1
Multípara	**Multípara = 2**	2
Sem informação	**Sem Informação = Em branco**	
Primigesta		1
Multípara		2
Multípara		2
Primigesta		1
Multípara		2
Multípara		2
Primigesta		1
Sem Informação		
Primigesta		1
Primigesta		1
Multípara		2
Sem Informação		

Se a variável for categórica, deve-se numerar com códigos suas categorias. Em caso de ausência de informação, pode-se deixar em branco.

Figura 8 – Exemplo de Variável Categórica com 1 resposta para cada paciente.

	ERRADO		CORRETO
paciente	Efeito Colateral		EfeitoColat
1	Cefaléia		1
2	Vômito	nome da variável	2
3	Nenhum	SEM espaço e Abreviada	
4	Cefaléia		1
5	Coceira	Efeito Colateral --> EfeitoColat	6
6	Náusea		5
7	Vômito		2
8	Nenhum		
9	Cefaléia	Codificar o Efeito Colateral	1
10	Coceira	1 = Cefaléia	6
11	Nenhum	2 = Vômito	
12	Vômito	3 = Taquicardia	2
13	Taquicardia	4 = Rubor Facial	3
14	Náusea	5 = Náusea	5
15	Taquicardia	6 = Coceira	3
16	Cefaléia	Em Branco = Nenhum	1
17	Nenhum		
18	Náusea		5
19	Taquicardia		3
20	Cefaléia		1

Se a variável for categórica com várias possibilidades de resposta, deve-se codificar com números cada uma delas, em caso de ausência de informação, pode-se deixar em branco.

No exemplo acima, a variável “Efeito colateral” tem 6 diferentes respostas, mas cada paciente tem apenas 1 efeito colateral, então cada resposta da variável foi codificada com um número (1= Cefaleia, 2= Vômito, ..., 6= Coceira).

Figura 9 – Exemplo de Variável Categórica com diversas respostas para cada paciente.

ERRADO

paciente	Efeito Colateral	Quais Efeitos Colaterais
1	Sim	Cefaléia,Náusea
2	Sim	Vômito,Taquicardia
3	Não	Nenhum
4	Sim	Cefaléia,Náusea,Taquicardia
5	Sim	Rubor Facial,Coceira,Náusea
6	Sim	Cefaléia,Náusea
7	Sim	Vômito,Taquicardia
8	Não	Nenhum
9	Sim	Cefaléia,Náusea,Taquicardia
10	Sim	Rubor Facial,Coceira,Náusea
11	Não	Nenhum
12	Sim	Vômito,Taquicardia
13	Sim	Taquicardia,Rubor Facial,Vômito
14	Sim	Cefaléia,Náusea,Taquicardia
15	Sim	Taquicardia,Rubor Facial,Vômito
16	Sim	Cefaléia,Náusea,Taquicardia
17	Não	Nenhum
18	Sim	Cefaléia,Náusea
19	Sim	Taquicardia,Rubor Facial,Vômito
20	Sim	Cefaléia,Náusea,Taquicardia

Criar 1 variável p/ cada Efeito Colateral

nome da variável SEM espaço e SEM acento

Codificar Resposta não ou ausente = 0 sim ou presente = 1

CORRETO

EfeitoColateral	Cefaleia	Nausea	Vomito	Taquicardia	RuborFacial	Coceira
1	1	1	0	0	0	0
1	0	0	1	1	0	0
0	0	0	0	0	0	0
1	1	1	0	1	0	0
1	0	1	0	0	1	1
1	1	1	0	0	0	0
1	0	0	1	1	0	0
0	0	0	0	0	0	0
1	1	1	0	1	0	0
1	0	1	0	0	1	1
0	0	0	0	0	0	0
1	0	0	1	1	0	0
1	0	0	1	1	1	0
1	1	1	0	1	0	0
1	0	0	1	1	1	0
1	1	1	0	1	0	0
0	0	0	0	0	0	0
1	1	1	0	0	0	0
1	0	0	1	1	1	0
1	1	1	0	1	0	0

Se uma determinada variável permitir múltiplas respostas, então ela deverá ser desmembrada em mais de uma variável, abrangendo as possíveis respostas para amostra estudada, conforme exemplo acima.

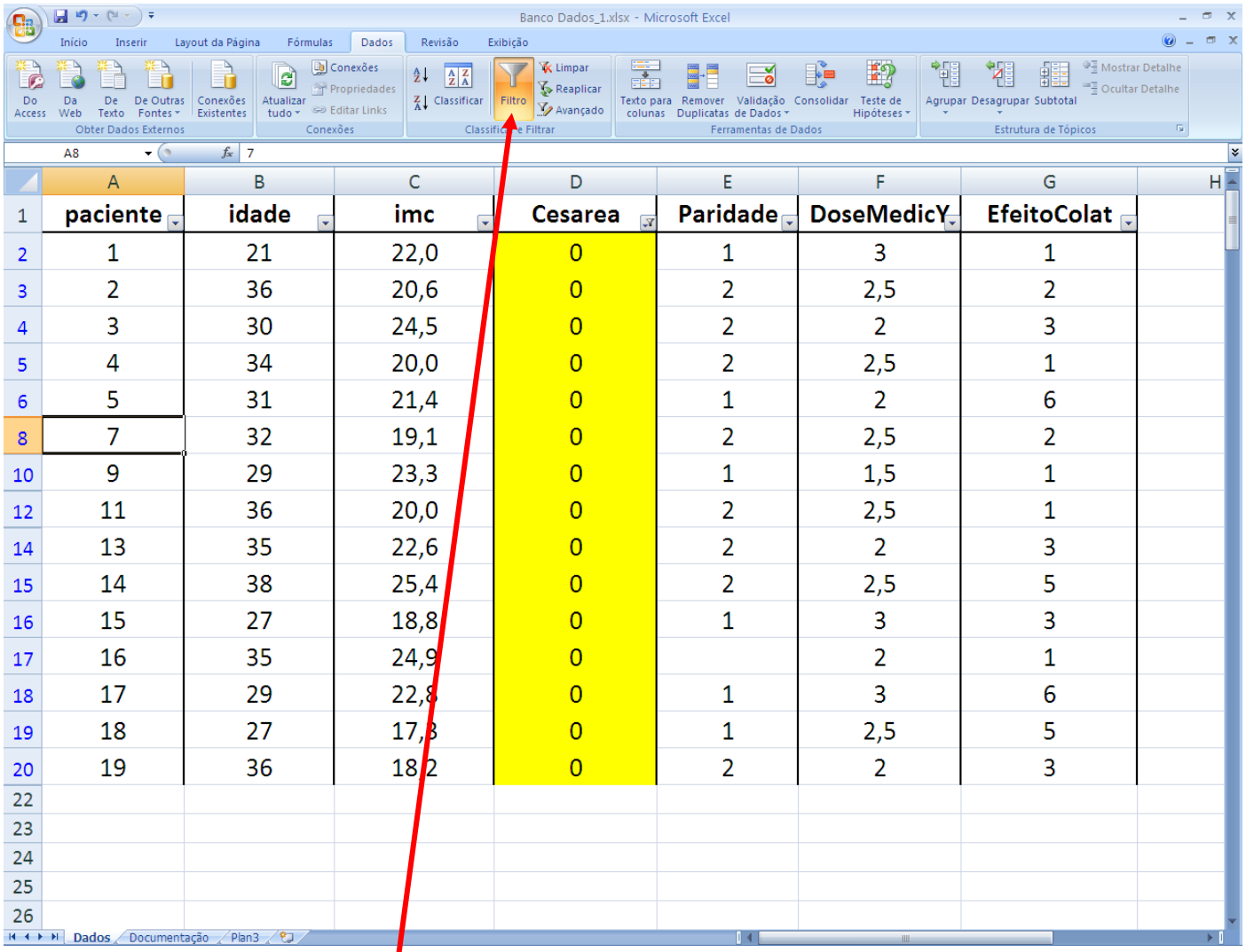
Na **Figura 9**, a variável “Quais Efeitos Colaterais” tem 6 (seis) tipos de efeitos (Cefaleia,Vômito,...,Coceira) , e cada paciente tem mais de um efeito colateral. Assim, para tornar fácil a inserção desse tipo de dado no software estatístico, a solução é criar 1 variável (coluna) para cada Efeito Colateral e codificá-lo como Ausente=0 e Presente=1.

O lado direito da Figura 9 mostra a Tabela Correta, onde a variável original “Quais Efeitos Colaterais” foi desmembrada em 6 variáveis (colunas), mostrando todos os 6 tipos de Efeitos Colaterais.

Ao percorrermos uma linha de dados, visualizamos todos os efeitos colaterais com seu status (ausente ou presente) de um determinado paciente.

5. Banco de Dados com opção de FILTRO.

Figura 10 – Banco de dados com Filtro mostra apenas os dados referentes à característica selecionada.



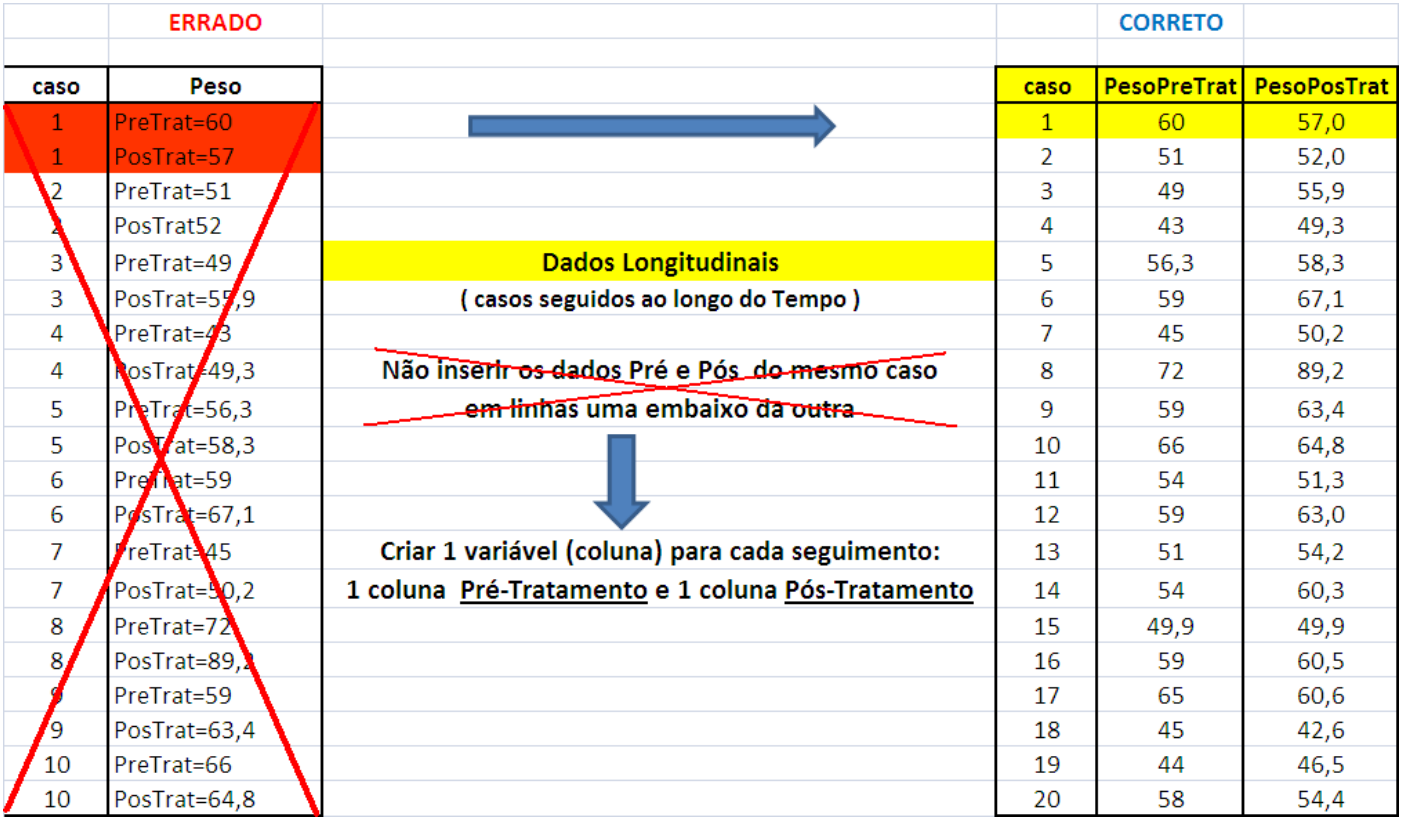
	A	B	C	D	E	F	G
1	paciente	idade	imc	Cesarea	Paridade	DoseMedicY	EfeitoColat
2	1	21	22,0	0	1	3	1
3	2	36	20,6	0	2	2,5	2
4	3	30	24,5	0	2	2	3
5	4	34	20,0	0	2	2,5	1
6	5	31	21,4	0	1	2	6
8	7	32	19,1	0	2	2,5	2
10	9	29	23,3	0	1	1,5	1
12	11	36	20,0	0	2	2,5	1
14	13	35	22,6	0	2	2	3
15	14	38	25,4	0	2	2,5	5
16	15	27	18,8	0	1	3	3
17	16	35	24,9	0		2	1
18	17	29	22,8	0	1	3	6
19	18	27	17,3	0	1	2,5	5
20	19	36	18,2	0	2	2	3
22							
23							
24							
25							
26							

A Tabela da **Figura 10** mostra apenas as pacientes que Não fizeram Cesárea (variável Cesárea = 0) e oculta (esconde) as pacientes que fizeram Cesárea. Note acima que as pacientes de número 6, 8,10,12 e 20 não aparecem, pois a opção **FILTRO** , esta ativada .

NUNCA utilizar a opção FILTRO, pois NÃO é possível visualizar todos os casos (pacientes) estudados, os dados ficam ocultos ocasionando diversos problemas para a análise descritiva e estatística dos dados.

6. Banco com dados Longitudinais.

Figura 11 – Banco de dados Longitudinais, Pré e Pós-tratamento,



Para estudos com dados Longitudinais (casos seguidos ao longo do Tempo), criar 1 variável (coluna) para cada seguimento (Pré e Pós) . Assim, teremos 1 coluna para Pré-Tratamento e 1 coluna para Pós-Tratamento. **Desta forma, visualizamos todos os dados de cada paciente em uma única linha da tabela**.

É de suma importância que todas as características do paciente estejam dispostas ao longo de apenas uma linha, isto facilita os procedimentos e cálculos que serão realizados no software Estatístico.

7. Banco com dados Hierárquicos.

Figura 12 – Banco com dados hierárquicos.

	ERRADO		CORRETO	
paciente	Pratica Ativ Física		PraticaAtivFisica	DiasPorSemana
1	Sim, 2 dias por semana		1	2
2	Não		0	0
3	Sim, 4 dias por semana		1	4
4	Sim, 1 dias por semana		1	1
5	Sim, 3 dias por semana	**Codificar**	1	3
6	Não	**PraticaAtivFisica**	0	0
7	Sim, 2 dias por semana	**Não = 0**	1	2
8	Sim, 1 dias por semana	**Sim = 1**	1	1
9	Sim, 3 dias por semana		1	3
10	Sim, 1 dias por semana		1	1
11	Sim, 4 dias por semana	**Criar variável**	1	4
12	Sim, 2 dias por semana	**DiasPorSemana**	1	2
13	Não	**(Quantidade de Dias por Semana**	0	0
14	Sim, 3 dias por semana	**que Pratica Ativ Física)**	1	3
15	Sim, 3 dias por semana		1	3
16	Sim, 2 dias por semana		1	2
17	Sim, 4 dias por semana		1	4
18	Sim, 1 dias por semana		1	1
19	Sim, 3 dias por semana		1	3
20	Sim, 2 dias por semana		1	2

Se uma determinada variável for do tipo hierárquica, que quantifica a frequência da variável de interesse, então ela deverá ser desmembrada em duas variáveis. Uma variável que informa se o evento ocorre ou não, (Ex: Pratica Atividade Física ? Sim ou Não) ,e outra variável que quantifica a frequência do evento (Ex: Quantos Dias por Semana ? 0,1,2,...)

8. Bibliografia

- HULLEY, S.B.; CUMMINGS, S.R.; BROWNER, W.S.; GRADY, D.G., NEWMAN, T.B. **Delineando a pesquisa clínica.** 4ª ed. Porto Alegre: Artmed, 2015. 386p.

- NAVARRO, F.C. **Excel 2013**: técnicas avançadas.Brasport: Rio de Janeiro, 2014. 306p.