# ME111 - Atividade 06

*Profa. Tatiana Benaglia*

*13/05/2020 - 1S2020*

## Análise Descritiva Bivariada

Nessa atividade, vocês irão praticar os conceitos de análise descritiva bivariada. Ao final dessa atividade, o aluno deverá:

- apresentar tabelas de contingência (frequências absolutas e relativas) para resumir dados de duas variáveis qualitativas
- identificar o tipo de gráfico apropriado para representar a associação entre duas variáveis de acordo com os seus tipos e ser capaz de produzir esses gráficos no R ou em algum software.

### Introdução

Vocês irão utilizar o mesmo conjunto de dados da Atividade 04, onde fizeram uma análise descritiva univariada. Porém, o objetivo agora é fazer uma análise descritiva bivariada, ou seja, tentar encontrar associações entre pares de variáveis.

Para relembrar, o Sistema de Monitoramento de Fatores de Risco Comportamental ("Behavioral Risk Factor Surveillance System", BRFSS) é um *survey* anual realizado por telefone com 350.000 pessoas nos Estados Unidos, desenvolvido para identificar fatores de risco na população adulta e relatar tendências emergentes na saúde. A página do BRFSS (http://www.cdc.gov/brfss) contém uma descrição completa desta pesquisa, incluindo as questões de pesquisa que motivaram o estudo e muitos resultados interessantes derivados dos dados.

Nós focaremos numa amostra aleatória de 20.000 pessoas do BRFSS conduzido no ano de 2000 e um subconjunto de 9 variáveis, descritas a seguir:

- `genhlth`: status da saúde geral (excellent, very good, good, fair e poor).
- `exerany`: se praticou exercício físico no últim mês (1-sim; 0-não)
- `hlthplan`: se tem alguma forma de plano de saúde (1-sim; 0-não)
- `smoke100`: se fumou pelo menos 100 cigarros durante sua vida inteira (1-sim; 0-não)
- `height`: altura (em polegadas)
- `weight`: peso (em pounds)
- `wtdesire`: peso desejado (em pounds)
- `age`: idade (em anos)
- `gender`: gênero (f-feminino, m-masculino)

#### Leitura dos Dados

Primeiramente, iremos importar os dados das 20.000 observações para o `R`. Depois de inicializar o RStudio, execute o seguinte comando:

```
source("http://www.openintro.org/stat/data/cdc.R")
```

### Questões

1. Considere as variáveis `genhlth` e `exerany`.

a) Qual o tipo dessas variáveis? Eles estão representados de maneira correta no data frame?

b) Apresente um gráfico de barras para a variável `genhlth` apenas.

c) Resuma os dados em uma tabela de contingência de frequências absolutas com os níveis de `exerany` nas linhas e os níveis de `genhlth` nas colunas. Inclua também os totais das linhas e colunas. Dica: use a função `table()` e `addmargins()`.

d) Na tabela acima, o que representam os números no interior da tabela? Qual a casela com maior número de pessoas? Qual a casela com o menor número de pessoas?

e) Construa uma tabela que apresente a distribuição da variável saúde (`genhlth`) condicional no fato de ter feito ou não exercício no último mês (`exerany`), ou seja, de proporções condicionais. Dica: `prop.table()`.

f) Faça um gráfico de barras que represente a tabela do item anterior. A distribuição da variável saúde muda quando condicionamos no fato de ter feito exercício ou não? Baseado no gráfico, você diria que existe associação entre as duas variáveis?

g) Aplique um teste de hipótese apropriado para verificar se existe associação entre `genhlth` e `exerany`. Dica: `chisq.test()`. Quais são as hipóteses sendo testadas e suas conclusões?

2. Considere as variáveis `smoke100` e `gender`.

a) Apresente uma tabela de contingência com as propoções condicionais dos níveis de `smoke100` para cada gênero.

b) Faça um gráfico que represente a tabela acima.

c) Aplique um teste de hipótese apropriado para verificar se existe associação entre `smoke100` e `gender`. Quais são as hipóteses sendo testadas e suas conclusões?

3. O Índice Massa Corporal (IMC)) é uma medida internacional usada para calcular se uma pessoa está no peso ideal e é calculado pela fórmula:

$$IMC = \frac{peso}{altura^2},$$

cuja unidade é expressa em $kg/m^2$, ou seja, o peso em kilogramas e a altura em metros.

a) Calcule o IMC e adicione essa coluna aos dados `cdc`. Observe que as variáveis peso (`weigth`) e altura (`height`) estão em *pounds* e *inches*, respectivamente.

b) Calcule a correlação entre as variáveis IMC e peso. O valor calculado está de acordo com o que você espera dessa correlação?

c) Faça um gráfico de dispersão para IMC e peso. Comente o gráfico.

4. Considere as variáveis idade (`age`) e saúde (`genhlth`).

a) Calcule as estatísticas sumárias de idade para cada nível de saúde.

b) Faça um boxplot de age pelos níveis de saúde. Comente esse gráfico.

5. Olhando sua atividade como um todo, comente fatos interessantes que observou através de sua análise.