Paradoxo de Simpson: uma abordagem em R

Douglas Cardoso

7/16/2021

Quando estamos fazendo uma análise de dados, quase sempre temos que tomar cuidado para não cair em algumas pegadinhas contidas nas amostras, que podem nos fazer achar e inferir algo que na verdade é outra coisa. Por conta disso, os queridos estatísticos sempre nos oferecem algum famoso *teste estatístico* para sabermos com algum nível de certeza certo aspecto dos nossos dados. Como exemplo, se os dados são normalizados ou não, temos quatro principais testes: D'Agostino-Pearson, Anderson-Darling, Shapiro-Wilk e Kolmogorov-Smirnov. Há diversas opções de testes para muitos elementos estatísticos, e isso serve essencialmente para evitarmos de cair em erro de forma fácil. Neste artigo, irei apresentar uma dessas "pegadinhas estatísticas", com uma pitada de programação em R e alguns gráficos bonitinhos. Vamos lá!

Para começar, irei apresentar a primeira base de dados que utilizaremos.

`Palmerpenguins`

A base de dados `palmerpenguins` é um pacote no R que contém dois *datasets* com dados reais, coletados em documentos oficiais, de três pinguins em três ilhas no Arquipélago de Palmer, na Antártica. Esses *datasets* contém dados sobre tamanho de bico, tamanho da asa e peso corporal dos pinguins, o que nos oferece boas amostras para fazermos gráficos e análises estatísticas, como um substituto da famosa base de dados `iris`.

Aqui, demonstro as cinco primeiras linhas de nosso *dataset*. Temos 8 colunas e dados numéricos e textuais.

```
palmerpenguins::penguins |>
  head() |>
  knitr::kable(align = "cccccccc")
```

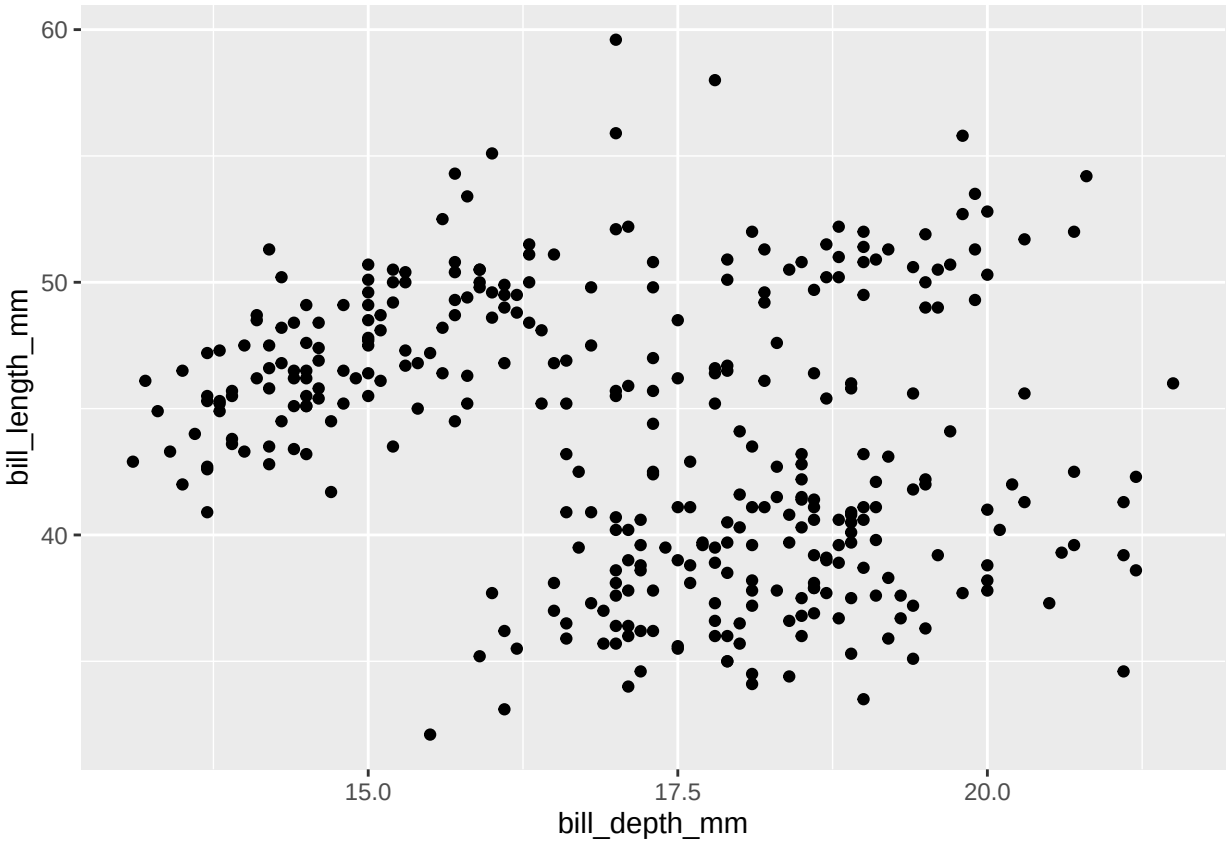
species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007

Indo direto ao ponto, temos a seguinte tarefa: **fazer um gráfico com a reta de regressão linear simples entre a profundidade do bico e o tamanho do bico de todos os pinguins na amostra**. Muito difícil? Vamos ver!

Utilizaremos as colunas `bill_depth_mm` e `bill_length_mm`, a primeira no eixo x e a segunda no eixo y. Para fazermos isso no R, utilizarei do operador pipe `|>` e da bibilioteca `ggplot2`, que utilizamos para fazer gráficos. A princípio, irei apenas plotar os dados em um gráfico de dispersão (aquele de pontinhos!).

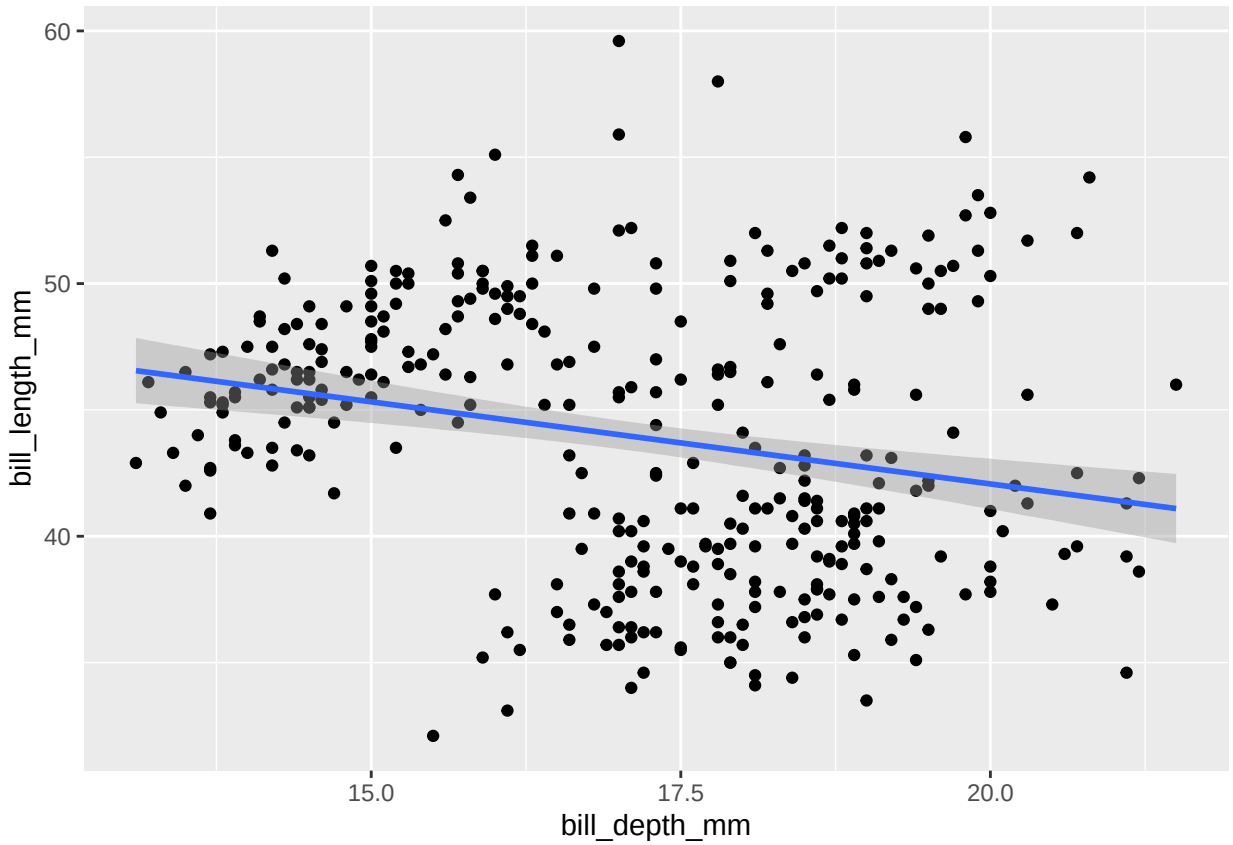
```
# Carregando as bibliotecas
library(palmerpenguins)
library(ggplot2)
```

```
penguins %>% # temos aqui nosso dataset
  ggplot(aes(x = bill_depth_mm, y = bill_length_mm)) + # selecionamos nossos eixos x e y
  geom_point() # selecionamos o tipo de gráfico que queremos, nesse caso, "de pontos", dispersão
```



Visualmente você, leitor, chutaria que os dados estão caminhando para qual direção? Cima? Baixo? Reto? Fica difícil né? Por isso temos métodos para entender essas relações sem ter que forçar nossa visão a tal ponto, para isso utilizaremos da regressão linear, que nos permite traçar uma linha reta no nosso gráfico capturando essa tendência.

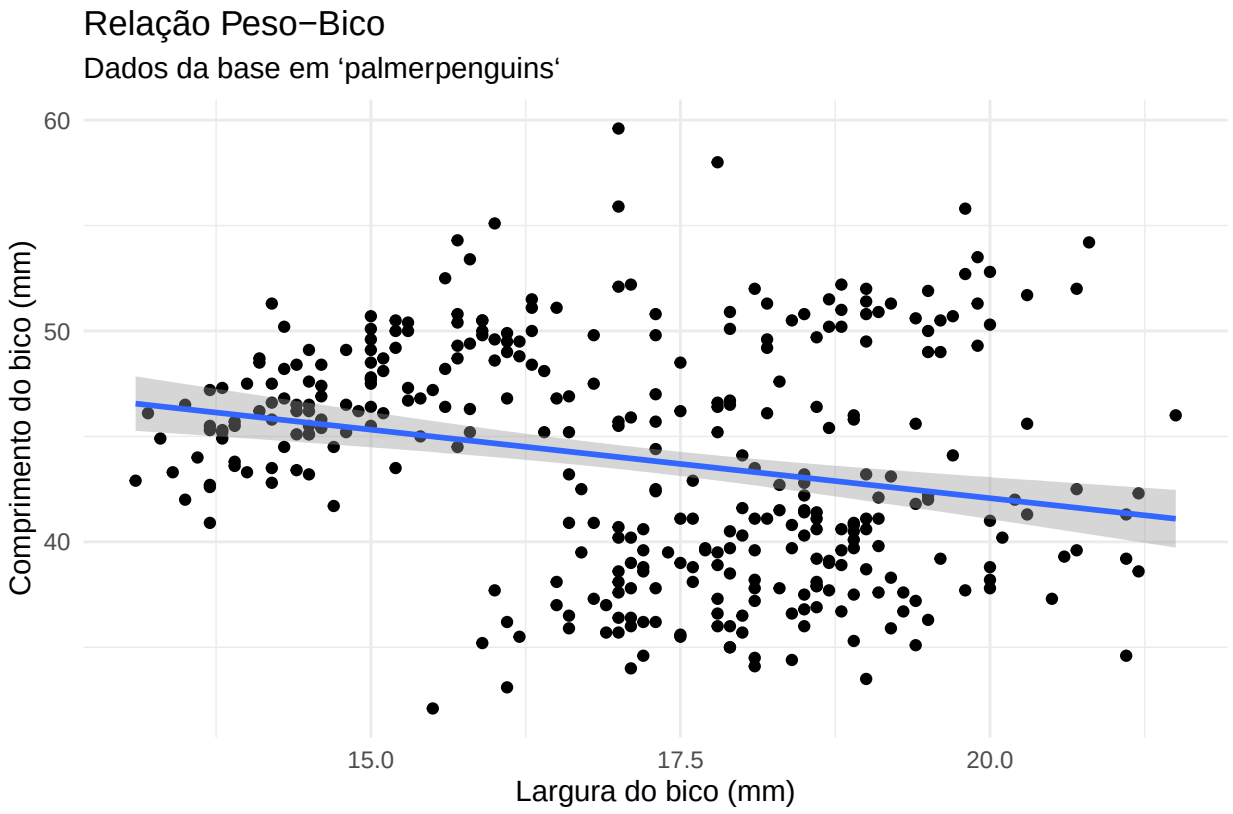
```
# Adicionamos o `geom_smooth` e colocamos como `method`, 'lm', que vem de 'linear model'
penguins %>%
  ggplot(aes(x = bill_depth_mm, y = bill_length_mm)) +
  geom_point() +
  geom_smooth(method = 'lm')
```



Eaí? Acertou a direção?! A interpretação que damos a esse gráfico é: a relação entre a profundidade do bico e o tamanho do bico é decrescente, isto é, quando maior `bill_depth_mm` menor é o valor de `bill_length_mm`, visto que a reta de regressão linear entre os dois tem inclinação para baixo, apesar da pouca intensidade. É basicamente isso que o gráfico mostra. "Estatisticamente falando", poderíamos dizer, pela análise visual, que essas duas variáveis possuem **correlação negativa entre si**.

Antes de partirmos para a pegadinha que nos aguarda, irei deixar nosso gráfico mais estiloso.

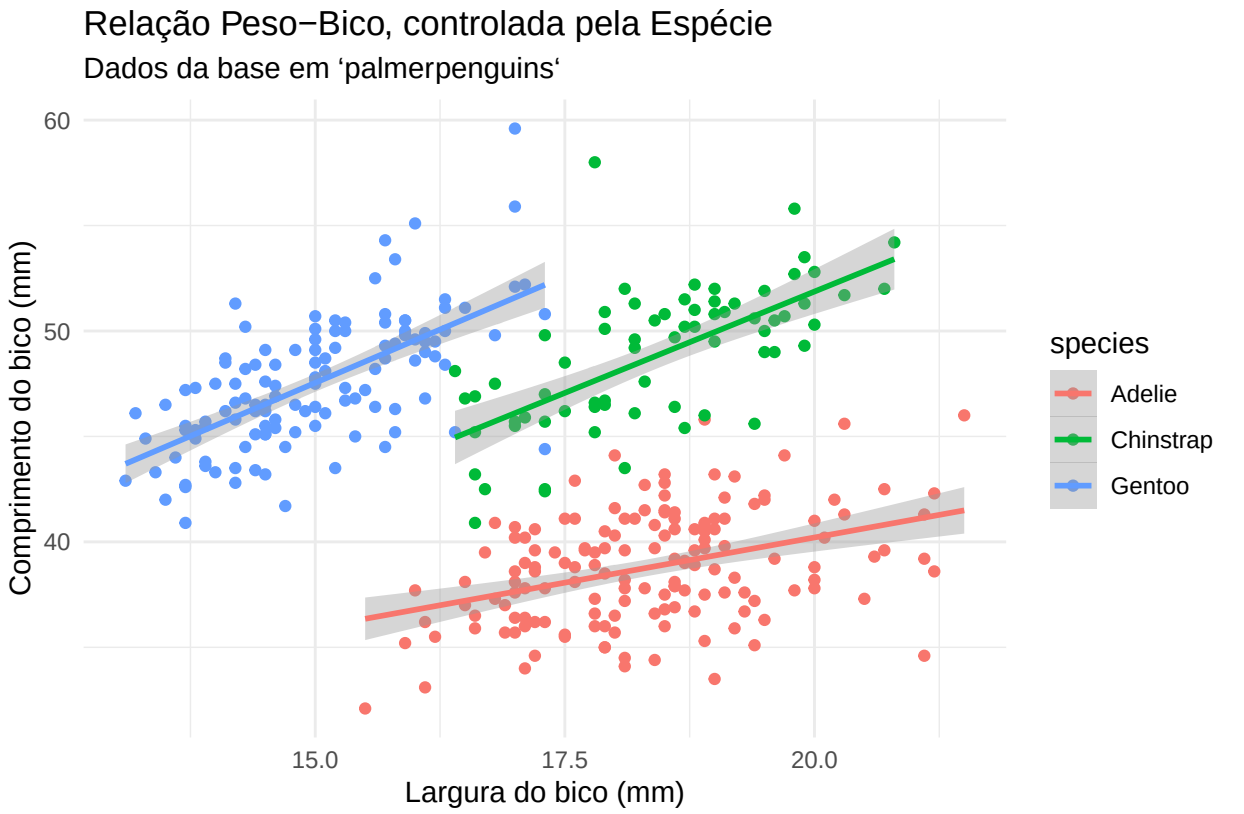
```
penguins %>%
  ggplot(aes(x = bill_depth_mm, y = bill_length_mm)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  theme_minimal() +              # escolha de um tema diferente para o gráfico
  labs(                          # mudando o nome dos eixos
    x = "Largura do bico (mm)",
    y = "Comprimento do bico (mm)",
    title = "Relação Peso-Bico",
    subtitle = "Dados da base em `palmerpenguins`",
    caption = "Fonte: Horst e Gorman (2020)")
```



Fonte: Horst e Gorman (2020)

Agora que sabemos a relação entre a largura e comprimento do bico entre todos os pinguins, seria interessante vermos essa mesma relação mas separado por espécie. Vamos ao R.

```
# Para discriminar por espécie, basta acrescentar o argumento `color = species`
penguins %>%
  ggplot(aes(x = bill_depth_mm, y = bill_length_mm, color = species)) +
  geom_point() +
  theme_minimal() +
  labs(
    x = "Largura do bico (mm)",
    y = "Comprimento do bico (mm)",
    title = "Relação Peso-Bico, controlada pela Espécie",
    subtitle = "Dados da base em `palmerpenguins`",
    caption = "Fonte: Horst e Gorman (2020)") +
  geom_smooth(method = 'lm')
```



Fonte: Horst e Gorman (2020)

Opa? O que aconteceu com a inclinação da reta? Inverteu! Recaptulando: quando plotamos a reta de regressão linear nos dados sem discriminação, ela se inclina para baixo, mas quando discriminamos por espécie, ela se inclina para cima, o exato contrário! Perceba que essa é uma pegadinha sagaz, com o primeiro gráfico poderíamos facilmente concluir que o tamanho e a profundida do bico possuem correlação negativa, e não teríamos porque desconfiar dessa afirmação, mas quando adicionamos um “detalhe”, a conclusão é outra. Esse é um exemplo do **Paradoxo de Simpson**:

> Se refere à existência de dados com associações estatísticas que se mantêm para a população total, porém se invertem para toda subpopulação

O paradoxo foi descrito pelo matemático e estatístico Edward Simpson em um *paper* em 1951, apesar de efeito similares serem descritos anteriormente por outros estatísticos. Há diversos exemplos envolvendo este paradoxo, e suas aplicações envolvem desde abstrações simples, como a que fizemos com o *dataset* dos pinguins à amostras sobre saúde. Mas, diante dessas conclusões, o que fazer para evitar ou resolver o paradoxo?

> Pensa causalmente na decisão entre ou segregar os dados em grupos ou agregá-los

“Pensar causalmente” é refletir em como os dados foram gerados e quais fatores influenciam seus resultados, incluindo os fatores que não constam em nossos dados. Perceba que não existe uma fórmula única para se decidir o que fazer. Uma das mais famosas aplicações do Paradoxo de Simpson é relacionado à saúde, o que carrega um peso de responsabilidade a mais por parte o analista, visto que envolve a saúde e muitas vezes a própria vida de seres humanos.

O que podemos aprender de pontos chaves nesse artigo é:

- Os dados por si só não são nadas
- Os dados que temos não nos mostram tudo, precisamos considerar o processo de geração dos dados
- Podemos ser facilmente enganados e para isso precisamos valorizar e utilizar dos testes estatísticos quando for conveniente
- *Think causally*

Abaixo deixo duas recomendações sobre o tema:

- Como o Paradoxo de Simpson explica estatísticas estranhas da COVID-19? (https://www.youtube.com/watch?v=t-Ci3FosqZs)
- Abordagem matemática e probablística do Paradoxo de Simpson: https://plato.stanford.edu/entries/paradox-simpson/