Estatística Descritiva

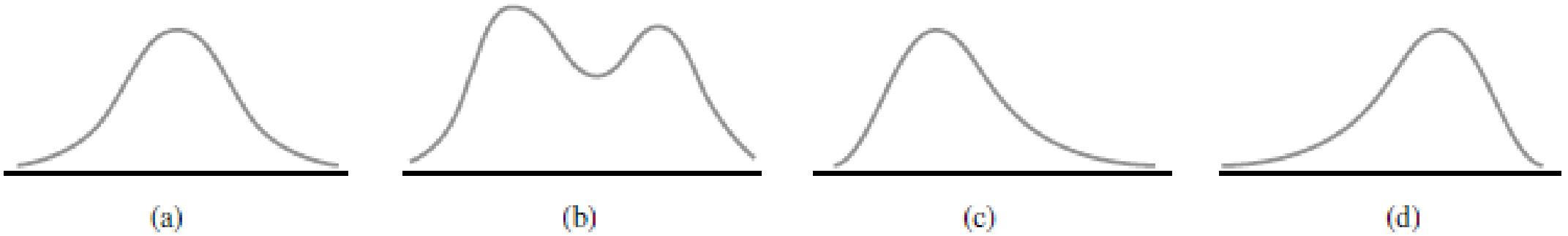
Ma. Pétala Tuy
Cientista de Dados – ATOS Brasil
Pesquisadora associada do Centro de Excelência em Pesquisa Aplicada
em Inteligência Artificial para a Industria do SENAI CIMATEC/ATOS

Conteúdo

- Simetria e Assimetria
- Transformações
- Análise Bidimensional

Simetria e Assimetria

- Os quantis podem ser úteis para se verificar se a distribuição dos dados é simétrica (ou aproximadamente simétrica).



(a) – Distribuição perfeitamente simetrica

(b) – Distribuição Bimodal

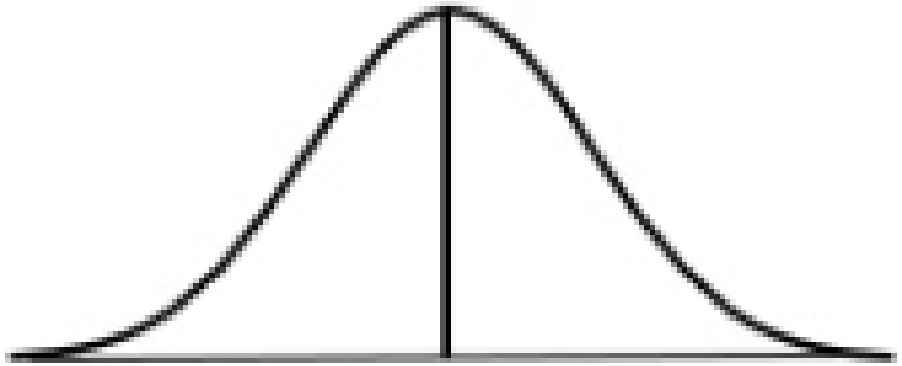
(c) – Distribuição com desvio positivo (assimétrica à direita)

(d) – Distribuição com desvio negativo (assimétrica à esquerda)

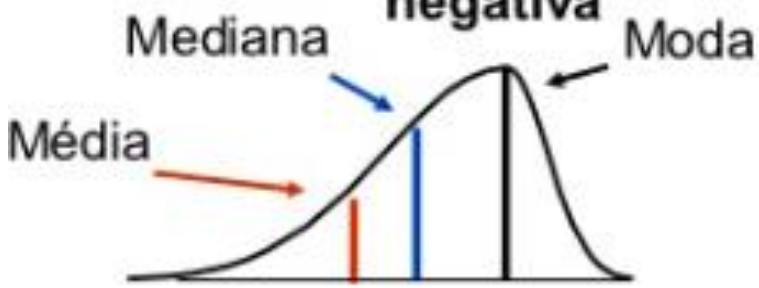
Simetria e Assimetria

Distribuição Simétrica

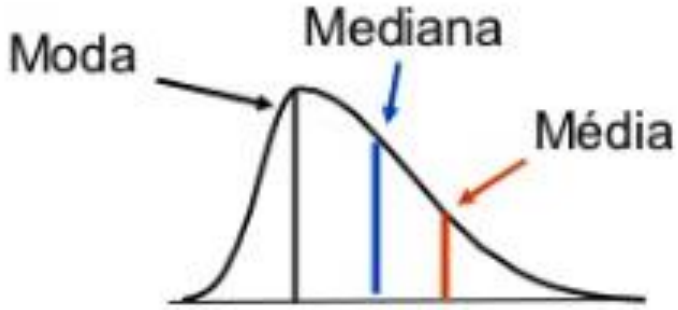
Média = Mediana = Moda



Assimetria à esquerda ou negativa



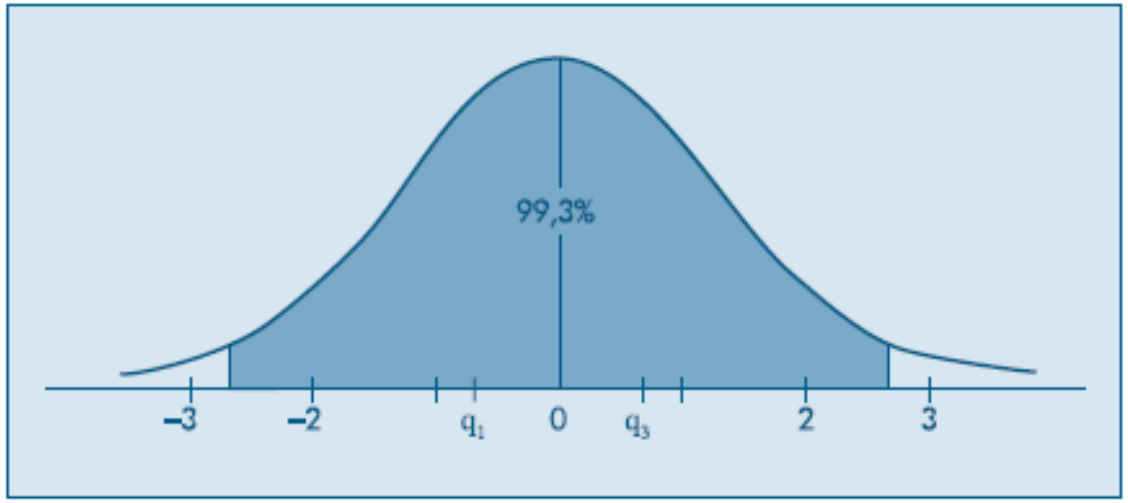
Assimetria à direita ou positiva



Simetria e Assimetria

- Porque checar assimetria?

Vários procedimentos estatísticos são baseados na suposição de que os dados provêm de uma Distribuição Normal (em forma de sino) ou então mais ou menos simétrica.



Transformações

- Para utilizar tais procedimentos estatísticos em distribuições assimétricas, deve-se efetuar uma transformação das observações, de modo a se obter uma distribuição mais simétrica e próxima da Normal.

- Uma família de transformações freqüentemente utilizada é

$$x^{(p)} = \begin{cases} x^p, & \text{se } p > 0 \\ \ell n(x), & \text{se } p = 0 \\ -x^p, & \text{se } p < 0. \end{cases}$$

Normalmente, experimenta-se valores de p na sequência:

–3, –2, –1, – 1/2, – 1/3, – 1/4, 0, 1/4, 1/3, 1/2, 1, 2, 3

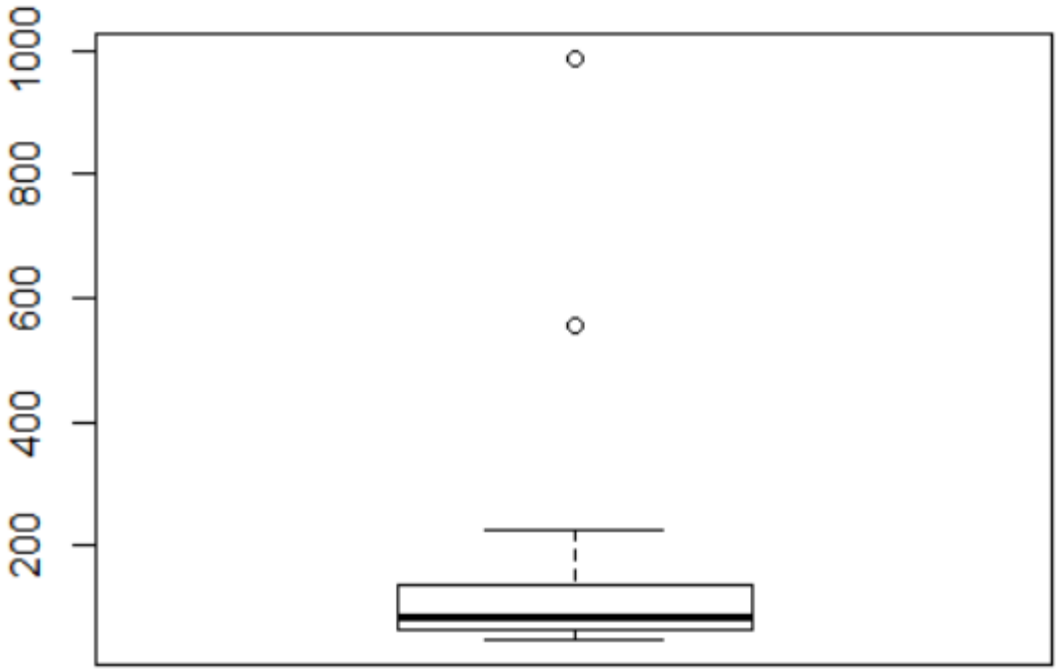
Transformações

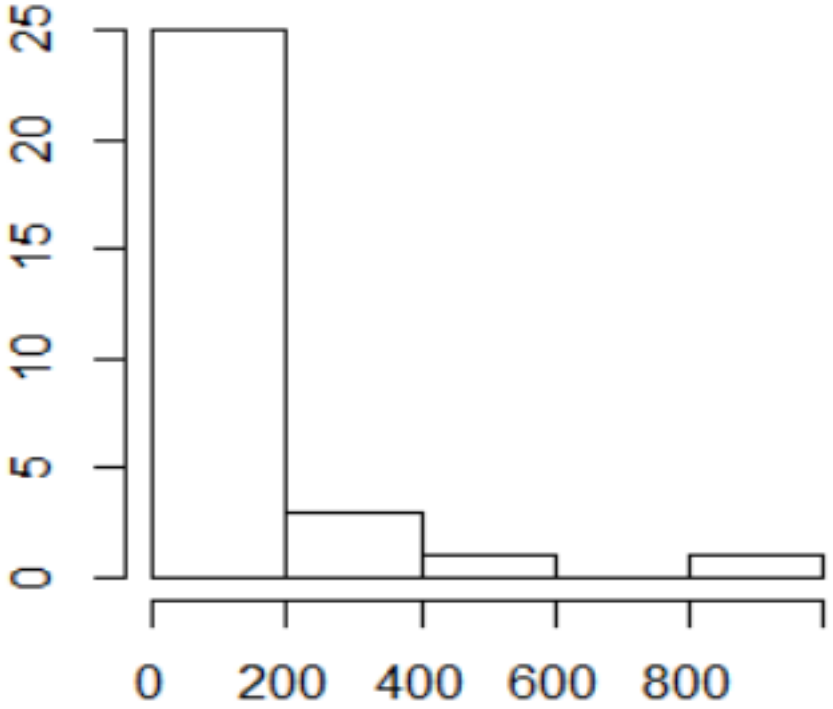
- Considere dados da população de alguns municípios do Brasil:

```
> head(dfpop)
        municipio populacao
1:      São Paulo      988.8
2: Rio de Janeiro      556.9
3:       Salvador      224.6
4: Belo Horizonte      210.9
5:      Fortaleza      201.5
6:       Brasília      187.0
```

Transformações

- Avaliando simetria da distribuição da variável “população”:





Transformações

Testando transformações:

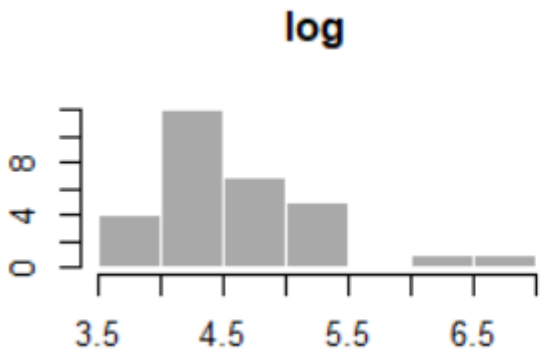
Log(x)

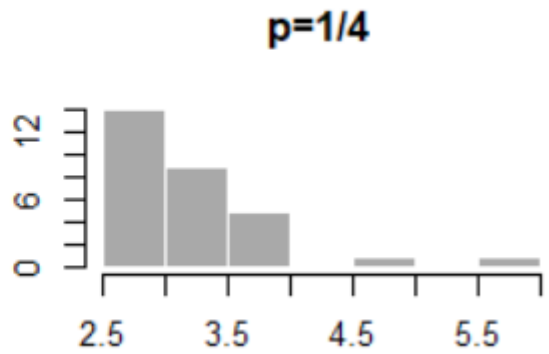
p = ¼

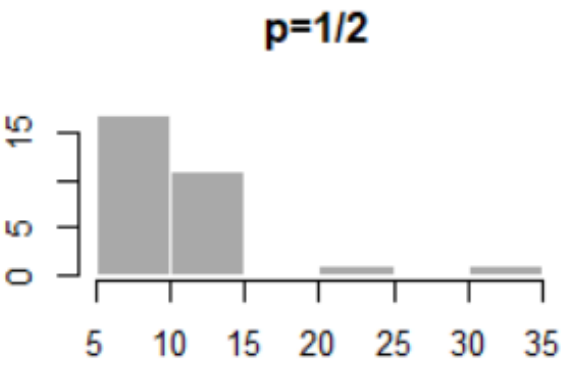
p = ½

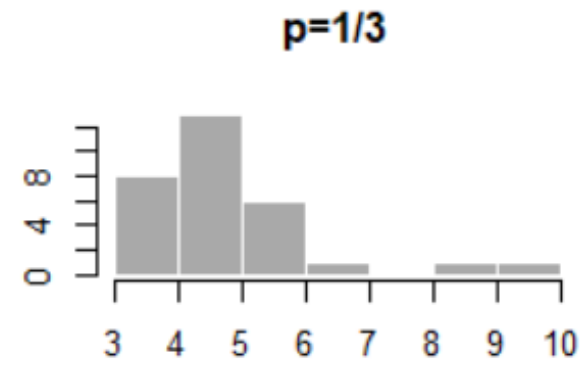
p = 1/3

Transformações **p=log** e **p=1/3** fornecem distribuições mais próximas de uma distribuição simétrica.









Análise Bidimensional

- Quando consideramos duas variáveis (ou dois conjuntos de dados), podemos ter três situações:

(a) as duas variáveis são qualitativas

(b) as duas variáveis são quantitativas

(c) uma variável é qualitativa e outra é quantitativa

- Em todas as situações, o objetivo é encontrar as possíveis relações ou associações entre as duas variáveis.

Análise Bidimensional

Duas variáveis qualitativas

- Quando as variáveis são qualitativas, os dados são resumidos em tabelas de dupla entrada (ou de contingência).

- Tabelas de contigência contém as freqüências absolutas ou contagens de indivíduos que pertencem simultaneamente a categorias de uma e outra variável.

```
> table(df$febre,df$classificacao_final)

      Confirmado Descartado Suspeito
NAO          720       2193      584
SIM          572        821      286
```

Análise Bidimensional

Duas variáveis qualitativas

- A distribuição do resultado final do teste difere entre os indivíduos que apresentaram febre e os que não apresentaram o sintoma?

Análise Bidimensional

Duas variáveis qualitativas

```
> tabFinal
             Confirmado Descartado Suspeito Total_coluna
NAO               13.91      42.37    11.28        67.56
SIM               11.05      15.86     5.53        32.44
Total_linha       24.96      58.23    16.81       100.00
```

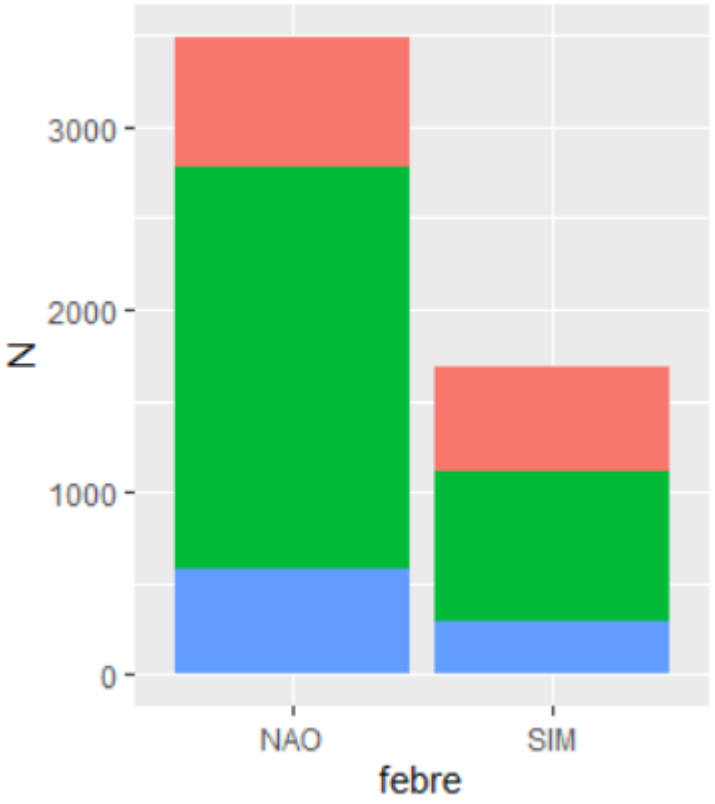
```
> (tabela <- round(prop.table(table(df$febre,df$classificacao_final),2),4)*100) # Por linha

    Confirmado Descartado Suspeito
NAO      55.73      72.76    67.13
SIM      44.27      27.24    32.87
> (tabela <- round(prop.table(table(df$febre,df$classificacao_final),1),4)*100) # Por coluna

    Confirmado Descartado Suspeito
NAO      20.59      62.71    16.70
SIM      34.07      48.90    17.03
```

Análise Bidimensional

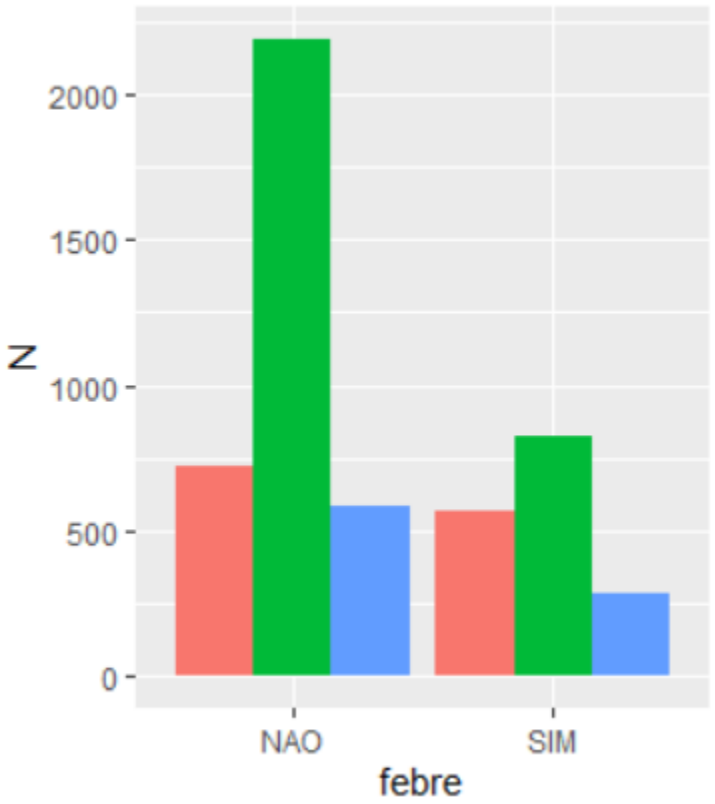
Duas variáveis qualitativas

- Análise gráfica para duas variáveis qualitativas.

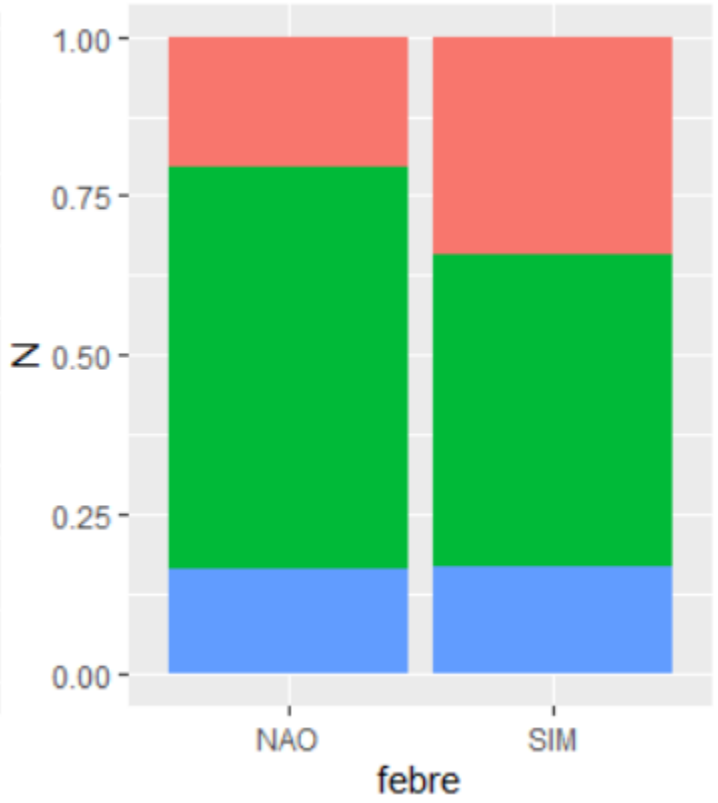
Opção (a)

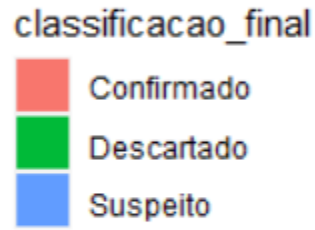


Opção (b)



Opção (c)





Análise Bidimensional

Duas variáveis qualitativas

- A distribuição do resultado final do teste difere entre os indivíduos que apresentaram febre e os que não apresentaram o sintoma?

Dentre os que apresentaram febre, houve uma maior proporção de casos confirmados em comparação aos que não apresentaram febre.

- Podemos dizer que a diferença é estatísticamente significante?

Análise Bidimensional

Duas variáveis qualitativas

Escolha duas variáveis qualitativas no dataset do COVID e realize análise bidimensional

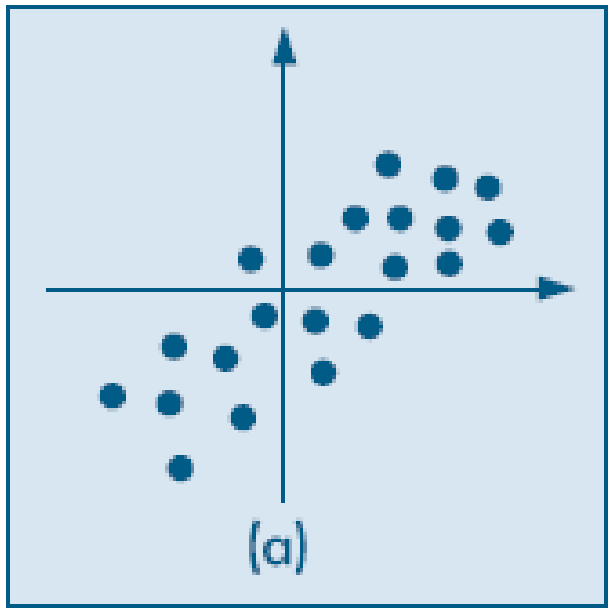
Análise Bidimensional

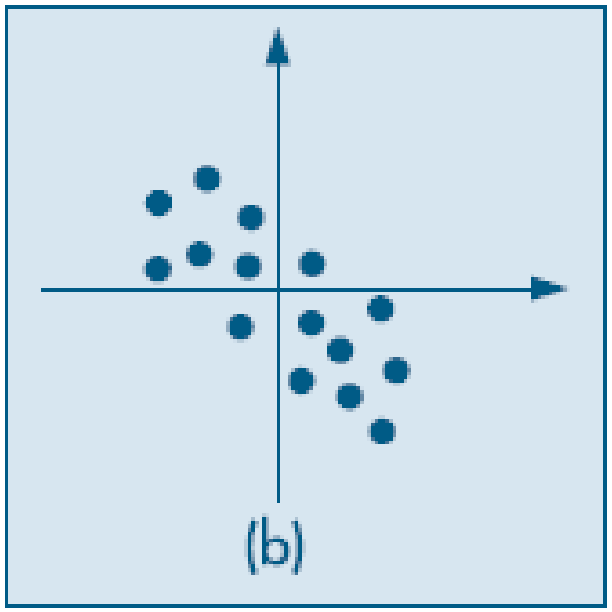
Duas variáveis quantitativas

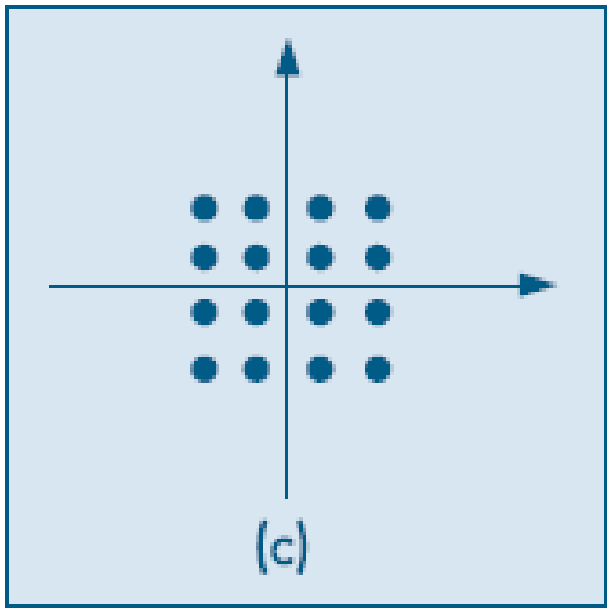
- Para análise bivariada entre duas variáveis quantitativas, pode-se:
    - Categorizar ambas as variáveis e utilizar as técnicas de análise para variáveis qualitativas.
    - Avaliar correlação entre as variáveis através do gráfico de dispersão

Análise Bidimensional

Duas variáveis quantitativas







(a) correlação linear positiva

(b) correlação linear negativa

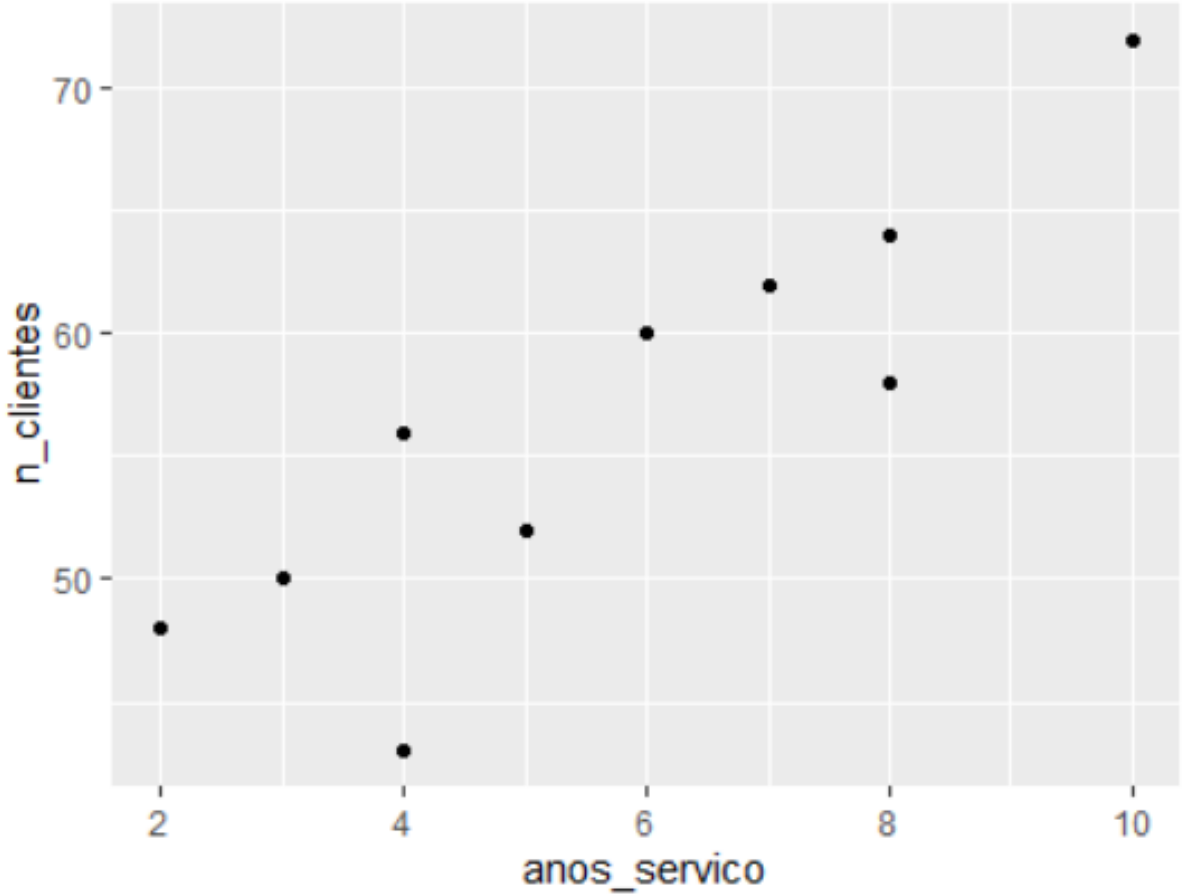
(c) não há associação linear entre as variáveis

Análise Bidimensional

Duas variáveis quantitativas

- Exemplo: Comparação entre número de clientes e anos de serviço

Qual tipo de correlação existe entre as variáveis?

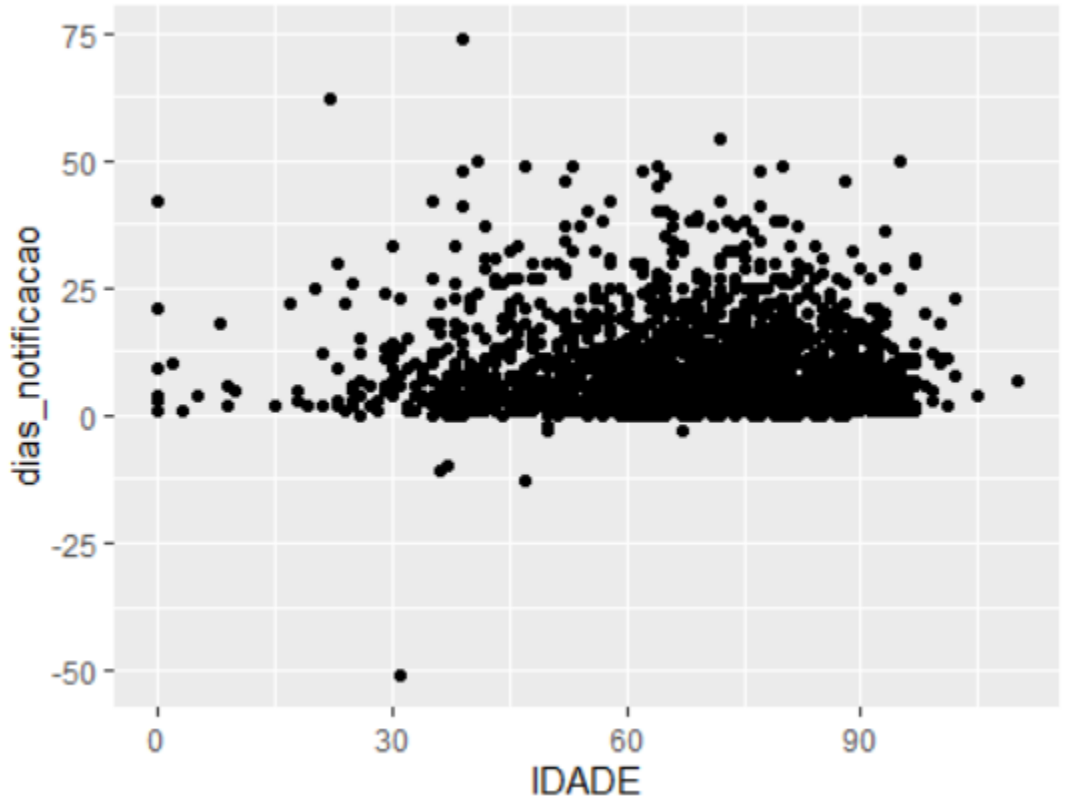


Análise Bidimensional

Duas variáveis quantitativas

- Existe correlação linear entre a idade do paciente que veio ao óbito e o número de dias até a notificação?

Faça uma análise descritiva univariada para a variável “dias até a notificação”.



Análise Bidimensional

Duas variáveis quantitativas

- O nível de associação entre as variáveis pode ser quantificado através do coeficiente de correlação linear:

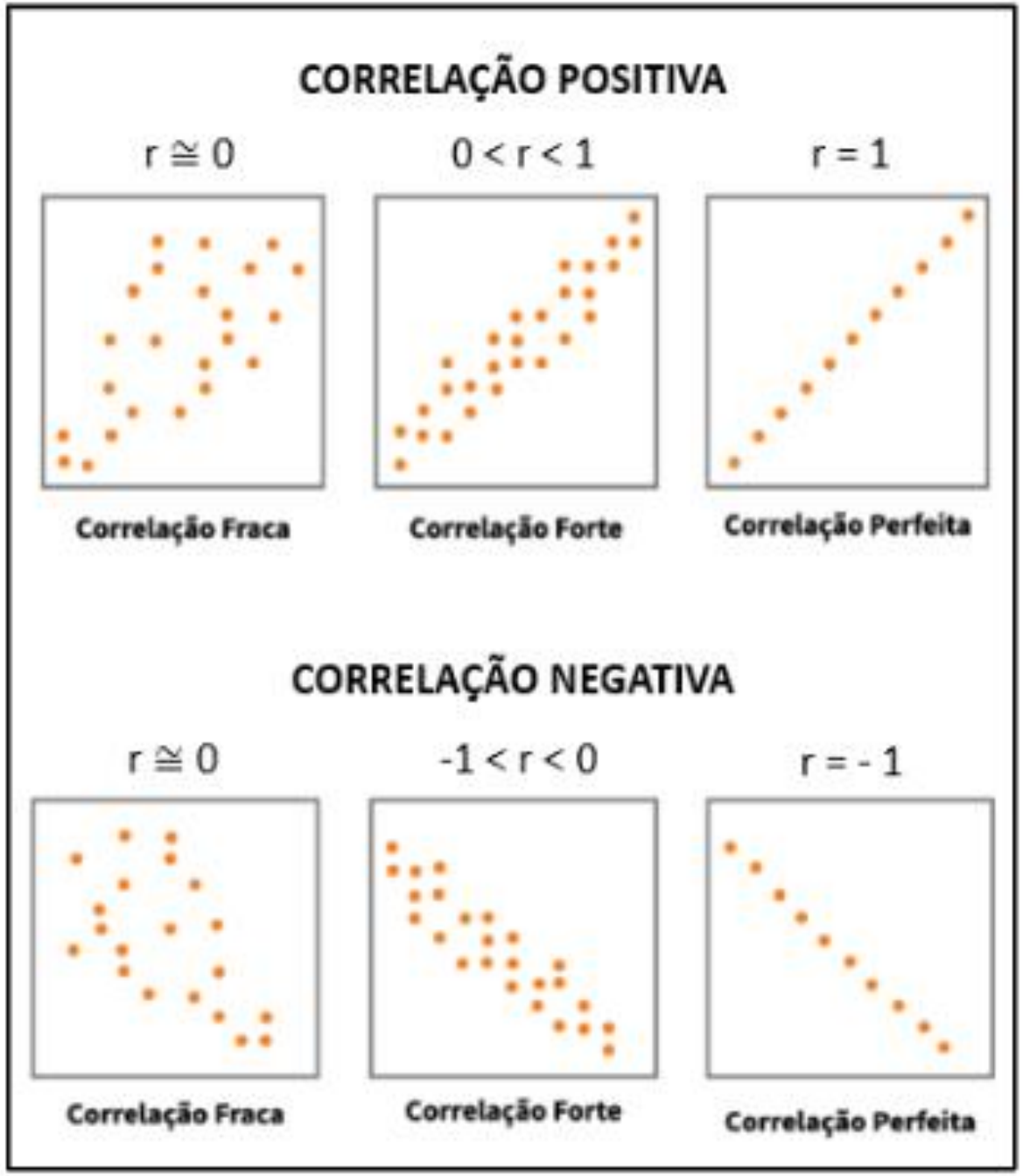
$$\operatorname{corr}(X,\ Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{dp(X)}\right)\left(\frac{y_i - \bar{y}}{dp(Y)}\right)$$

- O coeficiente de correlação, também definido por *r*, varia entre -1 e 1.

Análise Bidimensional

Duas variáveis quantitativas

- r ≈ 0
  - Não parece existir correlação entre as variáveis
- 0 < r < 1
  - Correlação linear positiva. Maiores valores de x levam a maiores valores de y
- -1 < r < 0
  - Correlação linear negativa. Maiores valores de x levam a menores valores de y
- r == -1 | r == 1
  - Correlação linear perfeita



Análise Bidimensional

Duas variáveis quantitativas

- Qual o nível de correlação entre a Idade em que o paciente veio à óbito e o número de dias até a notificação?

```
> cor(obitos$IDADE,as.numeric(obitos$dias_notificacao))
[1] -0.03389217
```

- Adicionar tabela de parâmetro de comparação

Análise Bidimensional

Variáveis qualitativa x quantitativa

- Para uma análise bidimensional entre uma variável quantitativa e uma variável qualitativa, é comum analisar o que acontece com a variável quantitativa dentro de cada categoria da variável qualitativa

- Essa análise pode ser conduzida por meio de:
  - Medidas-resumo de posição
  - Medidas-resumo de dispersão
  - Histogramas
  - Box plots

Análise Bidimensional

Variáveis qualitativa x quantitativa

- Exemplo:

Análise descritiva entre a variável qualitativa “Diabetes” e a variável quantitativa “Idade” para o dataset dos testados para o COVID-19.

```
> kable(df[!is.na(idade_anos) &
+       !is.na(diabetes) &
+       !diabetes %in% "IGNORADA",
+     list(mean=mean(idade_anos),
+          median = median(idade_anos),
+          sd=sd(idade_anos),
+          min = min(idade_anos),
+          max = max(idade_anos),
+          q1 = quantile(idade_anos,0.25),
+          q2 = quantile(idade_anos,0.50),
+          q3 = quantile(idade_anos,0.75),
+          n = .N),
+     by=diabetes])
```

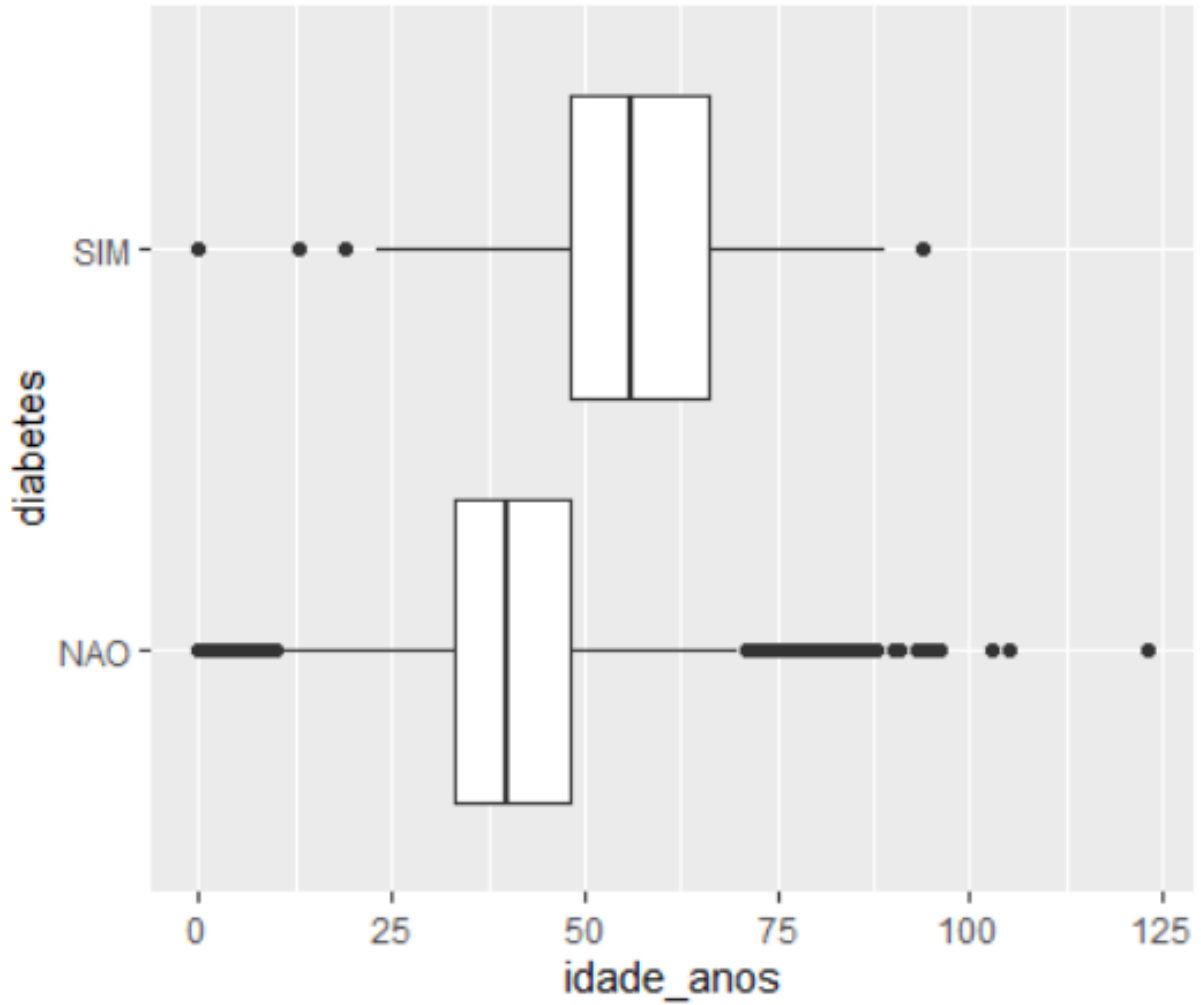
Escolha uma variável qualitativa de interesse a ser comparada com a variável Idade.

diabetes	mean	median	sd	min	max	q1	q2	q3	n
NAO	41.14271	40	13.32136	0	123	33	40	48	4912
SIM	56.74872	56	14.40123	0	94	48	56	66	195

Análise Bidimensional

Variáveis qualitativa x quantitativa

- A análise do boxplot sugere que:
  - A idade é maior para indivíduos que apresentam diabetes
  - A idade média para os que têm diabetes é 56, enquanto que a idade média para os que não possuem diabetes é 41
  - Parece haver uma maior variabilidade na idade para os que não têm diabetes

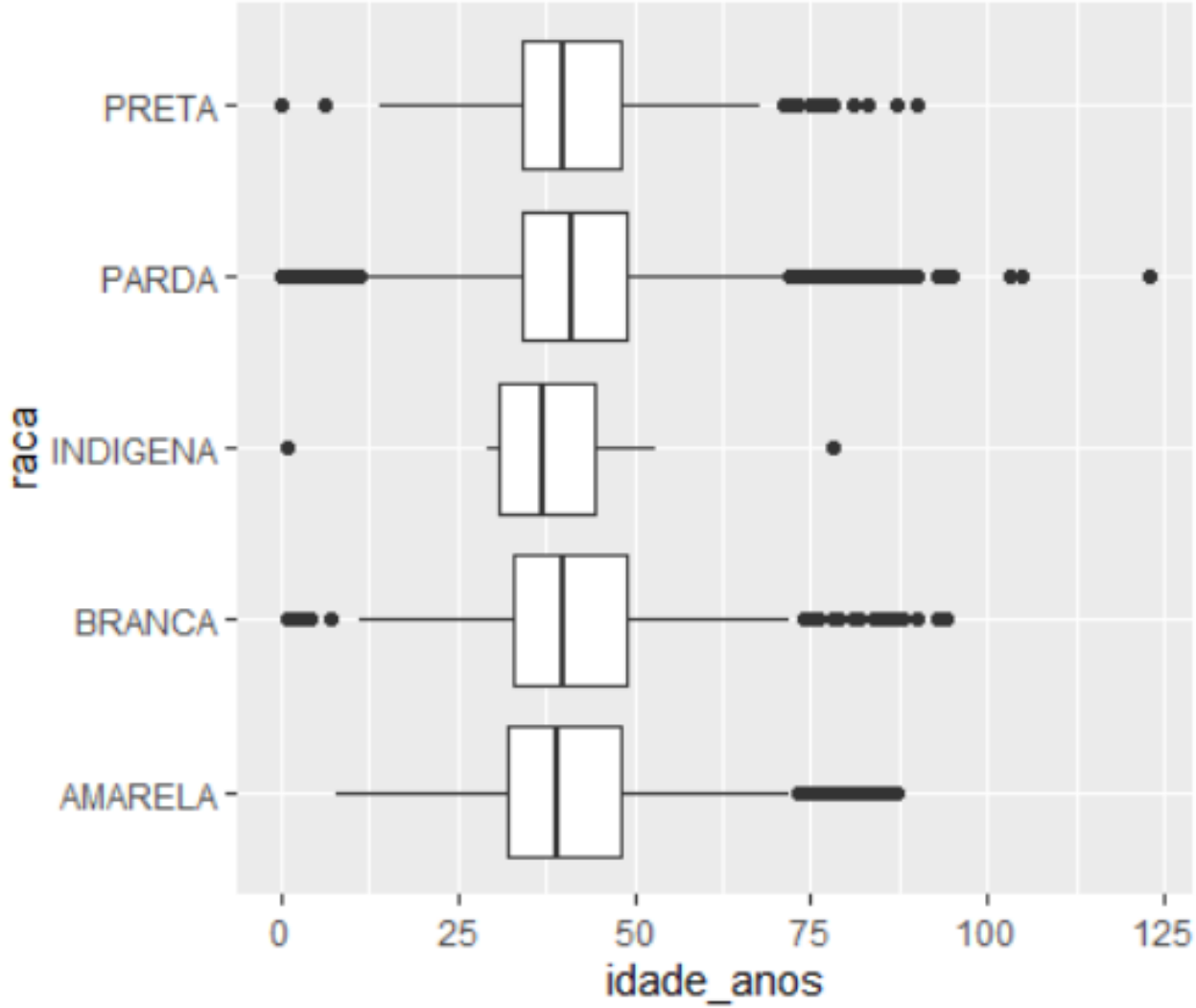


Análise Bidimensional

Variáveis qualitativa x quantitativa

- Não parece haver diferença na idade mediana entre os grupos de “Raça” nos testados pelo COVID-19.

- **Construa a análise gráfica entre a idade e a variável qualitativa escolhida no exercício anterior**



Análise Bidimensional

Variáveis qualitativa x quantitativa

- Como quantificar o grau de dependência entre as variáveis?
- As variâncias podem ser usadas para construir essa medida.
- A variância global da variável quantitativa é maior que as variâncias em cada categoria da variável qualitativa?

Variância Idade: **195,3**
Variância Idade entre diabéticos: **207,4**
Variância Idade entre não diabéticos: **192,4**
Variância média ponderada: **178,5**

$$\overline{\operatorname{var}(S)} = \frac{\sum_{i=1}^{k} n_i \operatorname{var}_i(S)}{\sum_{i=1}^{k} n_i},$$

Análise Bidimensional

Variáveis qualitativa x quantitativa

- Se a variância dentro de cada categoria for pequena e menor do que a global, significa que a variável qualitativa melhora a capacidade de previsão da quantitativa e portanto existe uma relação entre as duas variáveis.

- O grau de associação entre as duas variáveis pode ser medido como o ganho relativo na variância, obtido pela introdução da variável qualitativa.

0 < R2 < 1

$$R^2 = \frac{\mathrm{var}(S) - \overline{\mathrm{var}(S)}}{\mathrm{var}(S)} = 1 - \frac{\overline{\mathrm{var}(S)}}{\mathrm{var}(S)}$$

Análise Bidimensional

Variáveis qualitativa x quantitativa

- Para o exemplo de idade entre os diabéticos e não diabéticos:

```
> (r2 <- (varglobal-varponderada)/varglobal)
[1] 0.08568079
```

dizemos que 8,6% da variação total da idade é *explicada* pela variável "Diabetes".