

Correlação

Bioestatística em R

André M Ribeiro-dos-Santos

03 de 04, 2017

- Avaliar a associação entre medidas quantitativas.
- Reconhecer diferentes tipos de correlação.
- Ilustrar a relação entre medidas quantitativas.
- Reconhecer quando aplicar Pearson e Spearman.
- Conhecer principais transformações e quando aplicá-las.

Correlação

Em um estudo sobre diabetes, os pesquisadores observaram uma grande variação da sensibilidade à insulina entre os pacientes. Como trabalhos anteriores relacionaram essa variação com composição lipídica do tecido muscular. Foi medido a sensibilidade à insulina e composição de ácidos graxos de 10 pacientes.

A variação da sensibilidade à insulina está relacionada a composição de ácidos graxos?

Table 1: Medidas de sensibilidade à insulina e composição de ácido graxos em diabéticos

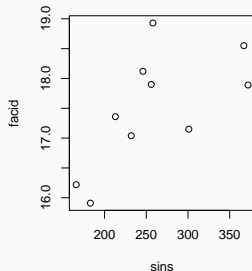
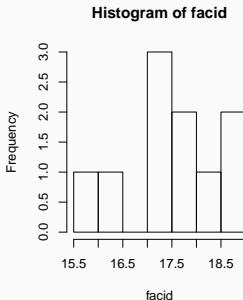
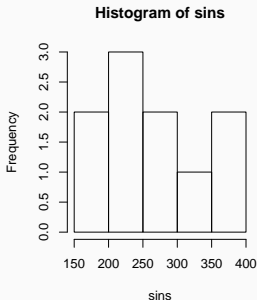
Sensibilidade à Insulina	Ácidos Graxos (%)	Sensibilidade à Insulina	Ácidos Graxos (%)
183	15.91	246	18.12
232	17.04	256	17.90
166	16.22	372	17.89
258	18.93	367	18.55
213	17.36	301	17.15

- As medidas em questão são categóricas ou quantitativas?
- Qual o tamanho da amostra?
- Qual a hipótese sendo avaliada?
- Qual a distribuição das medidas?

- As medidas em questão são categóricas ou quantitativas? **Ambas são quantitativas**
- Qual o tamanho da amostra? **10 pacientes**
- Qual a hipótese sendo avaliada? **As medidas são relacionadas.**

- Qual a distribuição das medidas? E como se relacionam?

```
> sins <- c(183, 232, 166, 258, 213, 246, 256, 372, 367, 301)
> facid <- c(15.91, 17.04, 16.22, 18.93, 17.36, 18.12, 17.90,
+           17.89, 18.55, 17.15)
> par(mfrow=c(1,3))
> hist(sins)
> hist(facid)
> plot(sins, facid)
```



Quando desejamos avaliar se a variação de uma medida afeta outra medida quantitativa, avaliamos a correlação linear das medidas com o coeficiente de correlação de Pearson (r).

$$r = \frac{cov_{xy}}{s_x * s_y} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \hat{x}) * \sum * y - \hat{y})}$$

Uma propriedade interessante deste coeficiente é que r^2 corresponde ao percentual da variabilidade em y explicada por x (ou vice-versa).

Valores do coeficiente

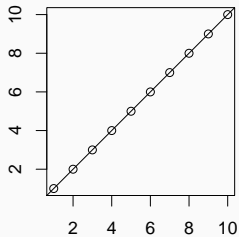
O *coeficiente de correlação* (r) assume valores entre -1 e 1, indicando uma correlação inversa em valores negativos, direta para valores positivos e zero quando não há correlação ¹.

Table 2: Interpretação dos valores do coeficiente de correlação.

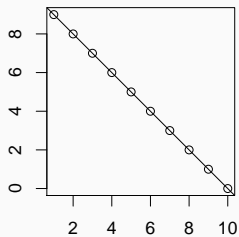
Coeficiente de Correlação	Interpretação
.90 to 1.00 (-.90 to -1.00)	Altíssima correlação
.70 to .90 (-.70 to -.90)	Alta correlação
.50 to .70 (-.50 to -.70)	Moderada correlação
.30 to .50 (-.30 to -.50)	Baixa correlação
.00 to .30 (.00 to -.30)	Praticamente nula

¹Mukaka M. A guide to appropriate use of Correlation coefficient in medical research. Malawi Medical Journal: The Journal of Medical Association of Malawi. 2012;24(3):69-71.

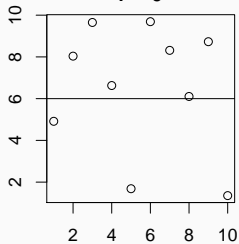
$r = 1$



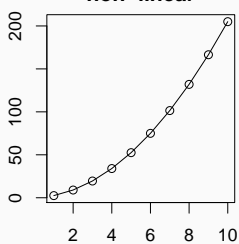
$r = -1$



$r = 0$



non-linear



```
> ?cor
```

```
> ## Correlation, Variance and Covariance (Matrices)
```

```
> ## Description:
```

```
> ##      'var', 'cov' and 'cor' compute the variance of 'x' and the
```

```
> ##      covariance or correlation of 'x' and 'y' if these are vectors.
```

```
> ##      If 'x' and 'y' are matrices then the covariances (or correlations)
```

```
> ##      between the columns of 'x' and the columns of 'y' are computed.
```

```
> ## Usage:
```

```
> ##      var(x, y = NULL, na.rm = FALSE, use)
```

```
> ##      cov(x, y = NULL, use = "everything",
```

```
> ##          method = c("pearson", "kendall", "spearman"))
```

```
> ##      cor(x, y = NULL, use = "everything",
```

```
> ##          method = c("pearson", "kendall", "spearman"))
```

```
> cov(sins, facid) /sqrt(var(sins) * var(facid))
```

```
## [1] 0.6468216
```

```
> cor(sins, facid)
```

```
## [1] 0.6468216
```

```
> cor(sins, facid)^2
```

```
## [1] 0.4183782
```

Teste de Correlação

Assumindo que r apresenta uma distribuição normal, podemos testar a significância da correlação observada.

$$H_0 : r = 0; \quad H_a : R \neq 0$$

```
> ?cor.test
> ## Test for Association/Correlation Between Paired Samples
> ## Description:
> ##      Test for association between paired samples, using one of
> ##      Pearson's product moment correlation coefficient, Kendall's
> ##      tau or Spearman's rho.
> ## Usage:
> ##      cor.test(x, y,
> ##              alternative = c("two.sided", "less", "greater"),
> ##              method = c("pearson", "kendall", "spearman"),
> ##              exact = NULL, conf.level = 0.95,
> ##              continuity = FALSE, ...)
```

```
> cor.test(sins, facid)
```

```
##
```

```
## Pearson's product-moment correlation
```

```
##
```

```
## data: sins and facid
```

```
## t = 2.3989, df = 8, p-value = 0.04325
```

```
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
```

```
## 0.02900965 0.90704747
```

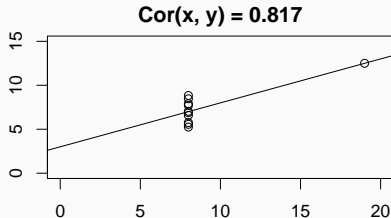
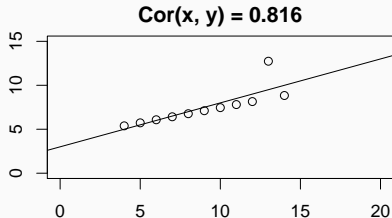
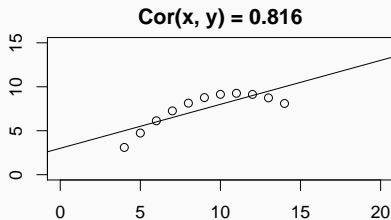
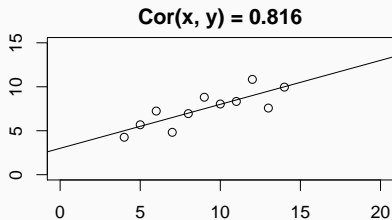
```
## sample estimates:
```

```
## cor
```

```
## 0.6468216
```

Erros comuns

1. Correlação não implica em causa.
2. Focar no P-value, no lugar do coeficiente.
3. Assumir correlação sem plotar relação.



Exercícios - Correlação de Pearson

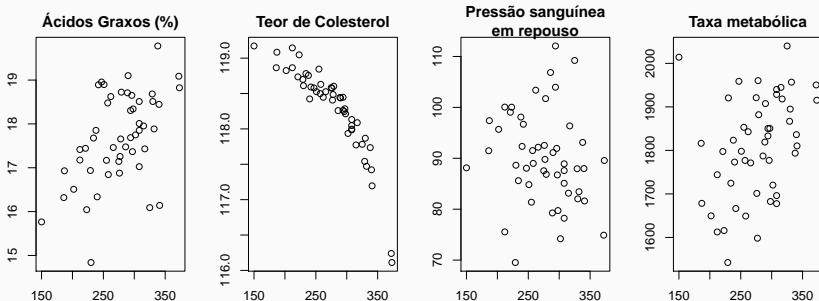
Pesquisadores buscando avaliar a associação entre diferentes medidas clínicas em pacientes de diabetes. Durante o estudo, eles coletaram os dados de 50 pacientes (disponíveis na tabela `db.tsv`). Sobre estes pacientes responda:

1. Ilustre a relação entre sensibilidade à insulina (`sen_ins`) com o percentual de ácido graxos (`fat_acid`), a taxa de colesterol (`chl`), pressão sanguínea em repouso (`bp`) e taxa metabólica (`m_rate`).
2. Alguma das relações a cima não poderia ser correlacionada por *Pearson*? Justifique
3. Calcule a correlação para qual *Pearson* se aplica, e indique quais possuem correlação significativa.
4. Reflita sobre o R^2 das correlações a cima e explique o seu valor.

```
> db <- read.table('db.tsv', header=T)
```


1. Ilustre a relação entre sensibilidade à insulina (`sen_ins`) com o percentual de ácido graxos (`fat_acid`), a taxa de colesterol (`chl`), pressão sanguínea em repouso (`bp`) e taxa metabólica (`m_rate`).

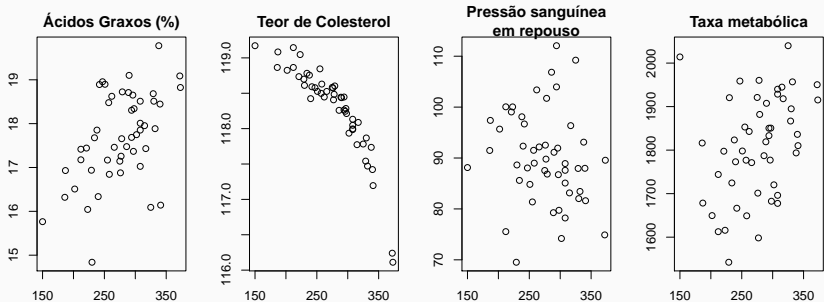
```
> par(mfrow = c(1,4), mar = c(2, 2, 2, 2))  
> plot(fat_acid~sen_ins, data=db, main="Ácidos Graxos (%)")  
> plot(chl~sen_ins, data=db, main="Teor de Colesterol")  
> plot(bp~sen_ins, data=db, main="Pressão sanguínea\nem repouso")  
> plot(m_rate~sen_ins, data=db, main="Taxa metabólica")
```



2. Alguma das relações a cima não poderia ser correlacionada por *Pearson*?

Justifique

Sensibilidade à insulina e o Teor de colesterol parece se relacionar de forma não linear, portanto não sendo recomendado utilizar a *correlação de Pearson*.



3. Calcule a correlação para qual *Pearson* se aplica, e indique quais possuem correlação significativa.

```
> with(db, cor.test(fat_acid, sen_ins))

##
## Pearson's product-moment correlation
##
## data: fat_acid and sen_ins
## t = 3.9634, df = 48, p-value = 0.0002447
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.2532044 0.6808056
## sample estimates:
## cor
## 0.4965555
```

```
> with(db, cor.test(bp, sen_ins))
```

```
##  
## Pearson's product-moment correlation  
##  
## data: bp and sen_ins  
## t = -1.1343, df = 48, p-value = 0.2623  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.4209813 0.1222812  
## sample estimates:  
## cor  
## -0.1615657
```

```
> with(db, cor.test(m_rate, sen_ins))
```

```
##
```

```
## Pearson's product-moment correlation
```

```
##
```

```
## data: m_rate and sen_ins
```

```
## t = 2.6166, df = 48, p-value = 0.01184
```

```
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
```

```
## 0.0831404 0.5751020
```

```
## sample estimates:
```

```
## cor
```

```
## 0.3533118
```

4. Reflita sobre o R^2 das correlações a cima e explique o seu valor.

```
> with(db, cor(fat_acid, sen_ins))^2
```

```
## [1] 0.2465673
```

```
> with(db, cor(bp, sen_ins))^2
```

```
## [1] 0.02610348
```

```
> with(db, cor(m_rate, sen_ins))^2
```

```
## [1] 0.1248292
```

Desejando estudar mais a fundo a relação entre sensibilidade à insulina e a composição lipídica dos pacientes de diabetes, eles decidiram investigar a concentração de colesterol no sangue. Na mesma amostra, os pesquisadores obtiveram os dados da concentração de colesterol em jejum.

A variação da sensibilidade à insulina está relacionada a concentração de colesterol no sangue?

Table 3: Medida de sensibilidade a insulina e colesterol em pacientes

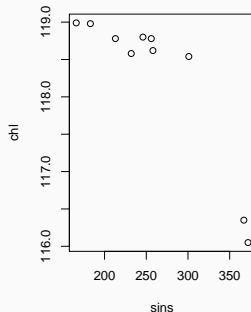
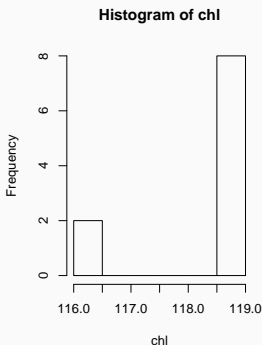
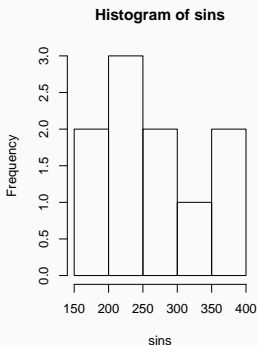
Sensibilidade à insulina	Colesterol	Sensibilidade à insulina	Colesterol
183	118.98	118.80	118.80
232	118.58	118.78	118.78
166	118.99	116.05	116.05
258	118.62	116.35	116.35
213	118.78	118.54	118.54

- As medidas em questão são categóricas ou quantitativas?
- Qual o tamanho da amostra?
- Qual a hipótese sendo avaliada?
- Qual a distribuição das medidas?

- As medidas em questão são categóricas ou quantitativas? **Ambas são quantitativas**
- Qual o tamanho da amostra? **10 pacientes**
- Qual a hipótese sendo avaliada? **As medidas são relacionadas.**

- Qual a distribuição das medidas?

```
> chl <- c(118.98, 118.58, 118.99, 118.62, 118.78, 118.80,  
+          118.78, 116.05, 116.35, 118.54)  
> par(mfrow = c(1, 3))  
> hist(sins)  
> hist(chl)  
> plot(sins, chl)
```

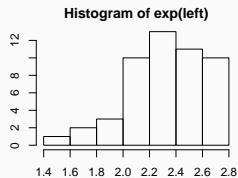
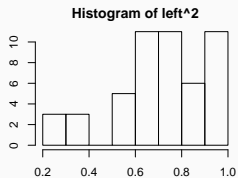
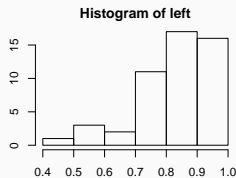
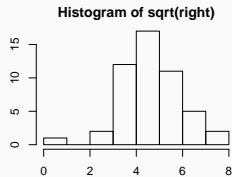
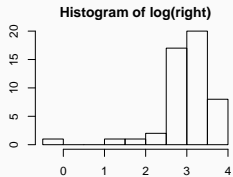
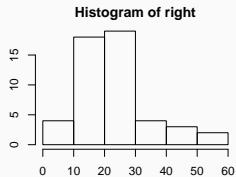
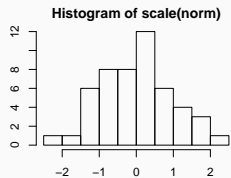
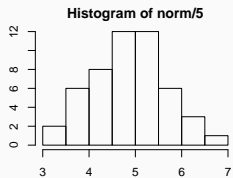
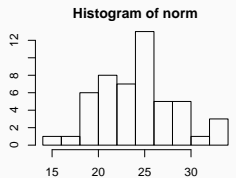


Quando os dados fogem a normalidade

1. Transformação

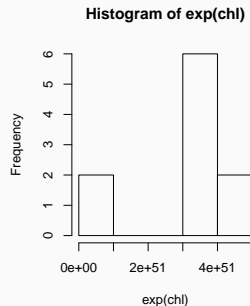
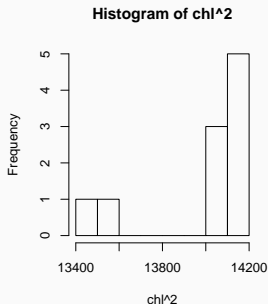
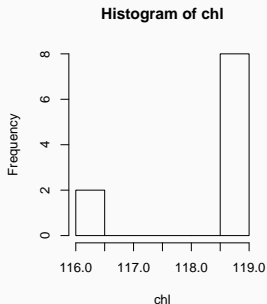
- **Raiz quadrada** (right skew, non-zero) `sqrt(x)`
- **Logaritmica** (right skew, non-zero) `log(x)`
- **Power** (left skew) x^2
- **Exponential** (left skew) `exp(x)`
- **Inverse** (conveniência) $1/x$ ou x^{-1}
- **Z-scale** (conveniência) `scale(x)`

2. Estatística Não-Paramétrica (*rank*)



Vamos, portanto aplicar uma transformação ao **colesterol (chl)** em busca de normalizar sua distribuição. Ela possui um forte *skew* para esquerda (*left skew*), podemos aplicar uma **potência** ou exponencial.

```
> par(mfrow = c(1,3))  
> hist(chl)  
> hist(chl^2)  
> hist(exp(chl))
```

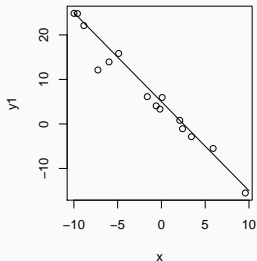


Correlação de Spearman

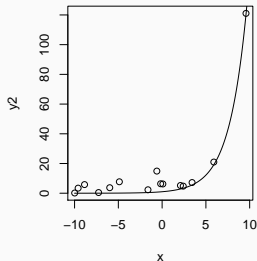
Quando não é possível corrigir o *skew* da amostra com uma transformação determinística, podemos recorrer à medidas não paramétricas como **rank**, a posição do valor quando todos os valores forem ordenados.

Numa correlação deseja-se associar o aumento de uma variável ao aumento ou decréscimo de outra, ou seja estabelecer uma **relação monotônica**.

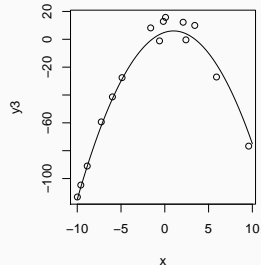
Monotônica



Monotônica

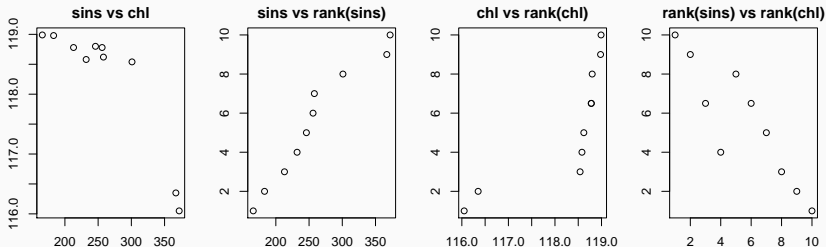


Não monotônica



No lugar de associar os valores reais, podemos-se trabalhar com o **rank**. A partir da relação entre o rank de ambas variáveis é calculada a **corelação de spearman**.

```
> par(mfrow=c(1,4), mar = c(2, 2, 2, 2))  
> plot(sins, chl, main = "sins vs chl")  
> plot(sins, rank(sins), main="sins vs rank(sins)")  
> plot(chl, rank(chl), main="chl vs rank(chl)")  
> plot(rank(sins), rank(chl), , main="rank(sins) vs rank(chl)")
```



```
> cor(sins, chl)
```

```
## [1] -0.8891366
```

```
> cor(sins, chl, method = "spearman")
```

```
## [1] -0.8875421
```

```
> cor.test(sins, chl, method="spearman")
```

```
##
```

```
## Spearman's rank correlation rho
```

```
##
```

```
## data: sins and chl
```

```
## S = 311.44, p-value = 0.0006097
```

```
## alternative hypothesis: true rho is not equal to 0
```

```
## sample estimates:
```

```
## rho
```

```
## -0.8875421
```

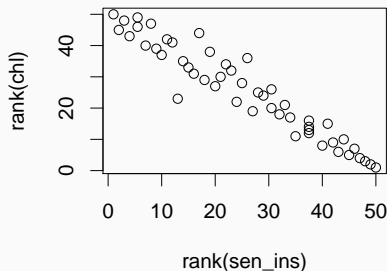
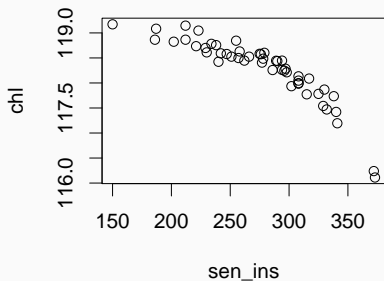

Exercícios - Correlação de Spearman

1. Previamente, os pesquisadores constataram que sensibilidade à insulina (`sen_ins`) e taxa de colesterol (`chl`) não apresentam uma relação linear. Mas existe alguma correlação entre as medidas? Ilustre a relação.
2. Tendo em vista que sensibilidade à insulina (`sen_ins`) possui uma forte correlação com o percentual de ácido graxos (`fat_acid`). Avalie se existe uma correlação entre o percentual de ácido graxos e a taxa de colesterol (`chl`).

```
> db <- read.table('db.tsv', header=T)
```

1. Previamente, os pesquisadores constataram que sensibilidade à insulina (**sen_ins**) e taxa de colesterol (**chl**) não apresentam uma relação linear. Mas existe alguma correlação entre as medidas? Ilustre a relação.

```
> par(mfrow = c(1,2))  
> plot(chl~sen_ins, data=db)  
> plot(rank(chl)~rank(sen_ins), data=db)
```



```
> with(db, cor.test(chl, sen_ins, method="spearman"))
```

```
##
```

```
## Spearman's rank correlation rho
```

```
##
```

```
## data: chl and sen_ins
```

```
## S = 40768, p-value < 2.2e-16
```

```
## alternative hypothesis: true rho is not equal to 0
```

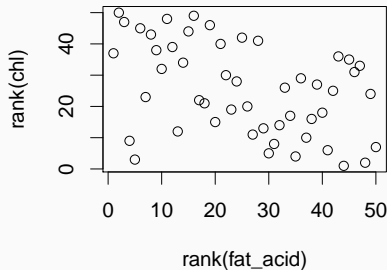
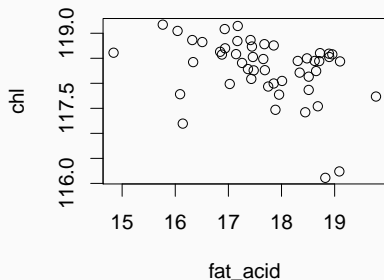
```
## sample estimates:
```

```
## rho
```

```
## -0.9576349
```

2. Tendo em vista que sensibilidade à insulina (**sen_ins**) possui uma forte correlação com o percentual de ácido graxos (**fat_acid**). Avalie se existe uma correlação entre o percentual de ácido graxos e a taxa de colesterol (**chl**).

```
> par(mfrow = c(1,2))  
> plot(chl~fat_acid, data=db)  
> plot(rank(chl)~rank(fat_acid), data=db)
```



```
> with(db, cor.test(chl, fat_acid, method="spearman"))
```

```
##
```

```
## Spearman's rank correlation rho
```

```
##
```

```
## data: chl and fat_acid
```

```
## S = 29762, p-value = 0.002047
```

```
## alternative hypothesis: true rho is not equal to 0
```

```
## sample estimates:
```

```
## rho
```

```
## -0.4291477
```

Ao final ...

- Avaliar a associação entre medidas quantitativas.
- Reconhecer diferentes tipos de correlação.
- Ilustrar a relação entre medidas quantitativas.
- Reconhecer quando aplicar Pearson e Spearman.
- Conhecer principais transformações e quando aplicá-las.

Até a próxima
