

Universidade Federal do Pará
Programa de Pós-graduação em Ciência Política
Disciplina: Tópicos Especiais Em Ciência Política:
Pesquisa Quantitativa em Ciência Política
Créditos: 4
Carga horária: 60h

AULA 9 CORRELAÇÃO

PROFA. NAIARA ALCANTARA E LUCAS OKADO



2023



Correlação Estatística: Pearson, Spearman e Kendall

1. Correlação de Pearson:

1. Avalia a relação linear entre duas variáveis contínuas.
2. Mede a força e a direção da relação linear.
3. Valores variam de -1 a 1.
 1. +1: Correlação perfeita positiva.
 2. -1: Correlação perfeita negativa.
 3. 0: Nenhuma correlação linear.
4. Supõe distribuição normal e linearidade.

2. Correlação de Spearman:

1. Avalia relação monotônica (não necessariamente linear) entre variáveis.
2. Adequada para dados contínuos ou ordinais.
3. Mede a força e direção da relação monotônica.
4. Usa classificações das observações, menos sensível a outliers.
5. Não faz suposições sobre a distribuição.

3. Correlação de Kendall (Tau):

1. Avalia concordância entre variáveis.
2. Adequada para dados contínuos ou ordinais.
3. Mede força e direção da relação de concordância.
4. Calculada a partir de concordâncias e discordâncias entre pares de observações.
5. Não faz suposições sobre a distribuição.

A escolha entre essas correlações depende dos dados e da relação que você deseja avaliar.

Pearson é para relações lineares, Spearman e Kendall são alternativas para relações não lineares ou quando não há suposições sobre a distribuição.

Spearman tende a não considerar os outliers, portanto é bastante provável que os resultados de Kendall sejam mais robustos

CORRELAÇÃO

- Os testes de correlação permitem indicar o quanto de X afeta Y e em qual direção este efeito acontece.

Desta forma uma medida de correlação indica a existência de uma relação entre duas variáveis, a intensidade desta relação e sua direção



CORRELAÇÃO

As correlações podem ser:

1. Paramétricas: quando envolvem duas variáveis contínuas com distribuição normal. É denominada de correlação de momentos de Pearson ou simplesmente de correlação de Pearson.
2. Não Paramétricas: quando são empregadas duas variáveis ordinais:
 1. Rho de Spearman ou correlação de Spearman
 2. Gamma de Goodman e Kruskal, ou gamma
 3. Tau B ou Tau C de Kendall

INTENSIDADE

- Uma correlação apresenta um valor que varia entre -1 e 1.
- Quanto mais distante de 0 (tanto para mais, quanto para menos) mais forte ela é.
- Desta forma, se plotarmos os valores de X e Y em um plano cartesiano (gráfico de dispersão), uma correlação perfeita ($r = 1,0$) apresentaria todos os seus pontos alinhados em uma mesma reta.
- O coeficiente é adimensional: uma correlação de 0,2 não é a metade de uma correlação de 0,4 ou um coeficiente de 0,3 não corresponde a 30%.

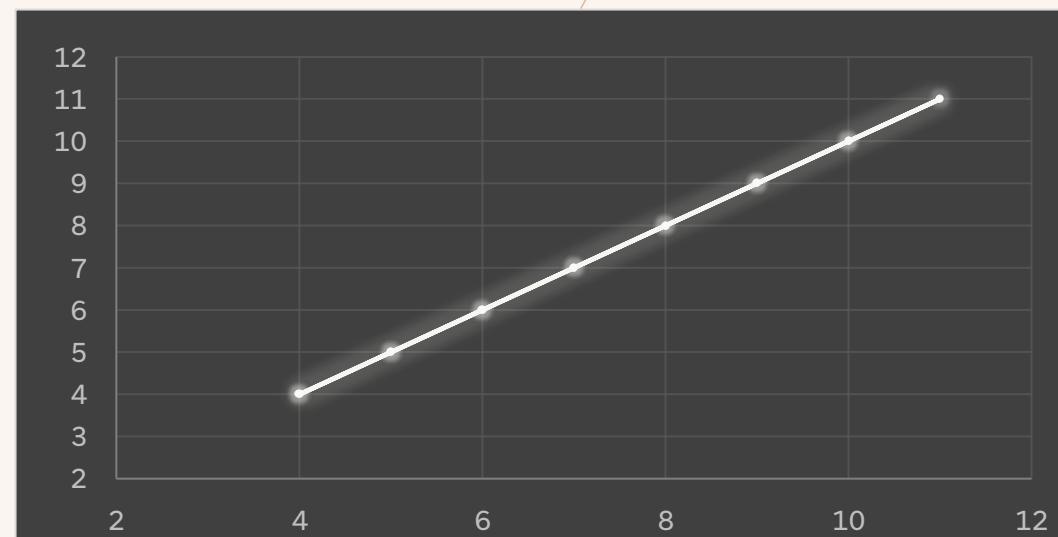
DIREÇÃO E INTENSIDADE

Além da intensidade, a correlação também indica a direção da relação de X e Y.

Esta informação é dada pelo sinal do valor da correlação:

Uma correlação positiva indica que um alto valor em X também ocasiona um valor alto em Y.

Já uma correlação negativa aponta que um valor em X causa um valor baixo em Y.

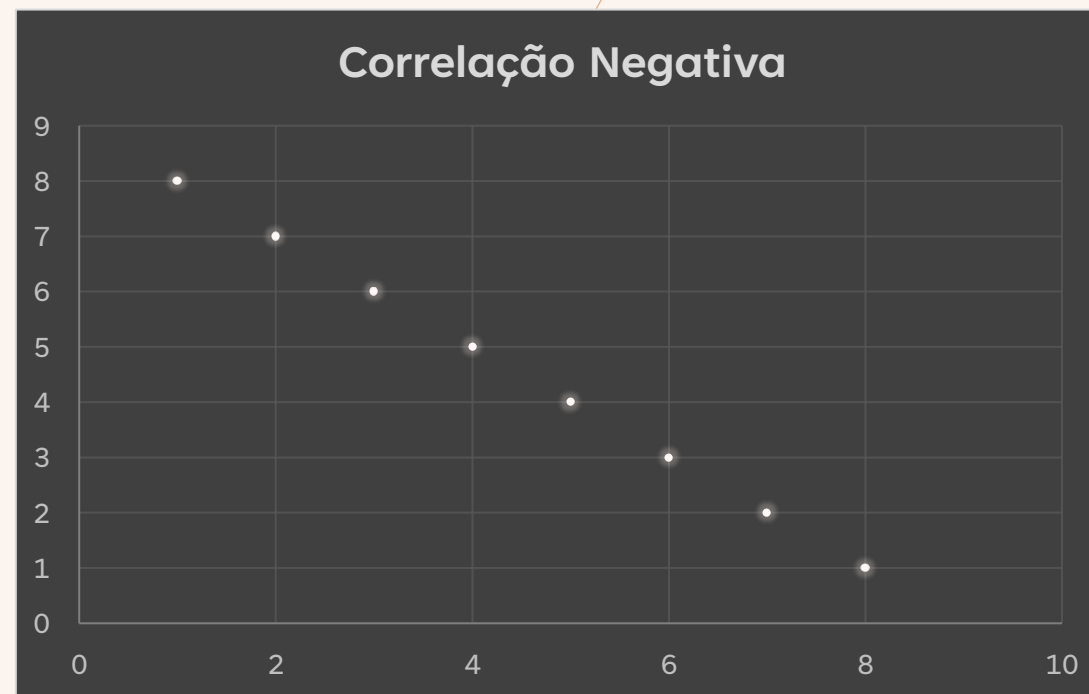
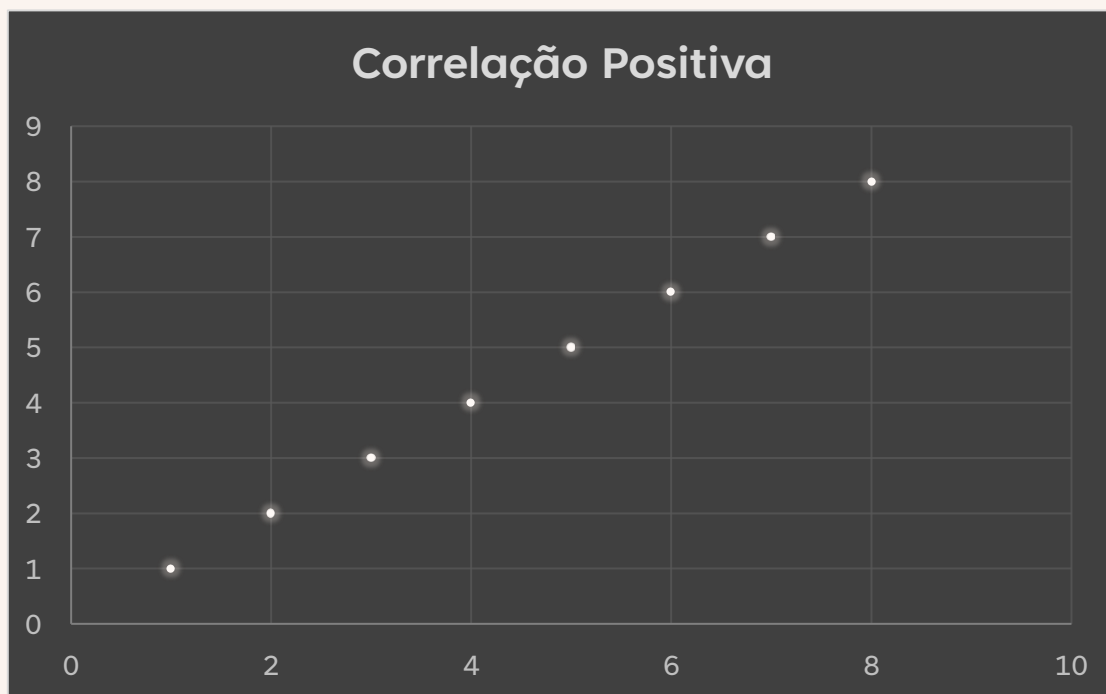


DIREÇÃO

Em um gráfico de dispersão os casos ficariam dispostos da seguinte forma:

Correlação positiva: do canto inferior esquerdo até o canto superior direito.

Correlação negativa: do canto superior esquerdo até o canto inferior direito.



CAUSALIDADE

- A existência de uma correlação entre duas variáveis não implica uma relação de causalidade entre elas. Ou seja, não é um indicativo de que X produziu um efeito em Y.
- Na ciência chamamos este fenômeno de Correlação Espúria. Isto acontece quando há uma correlação entre duas variáveis mas a inexistência de um nexos de causalidade entre elas.

<https://www.tylervigen.com/spurious-correlations>



CAUSALIDADE

A estatística é uma ferramenta que vai nos indicar a existência da relação entre duas ou mais variáveis.

Como cientistas devemos explicar o porquê deste fenômeno existir.

O que nos interessa é explicar o mecanismo causal da relação entre X e Y.

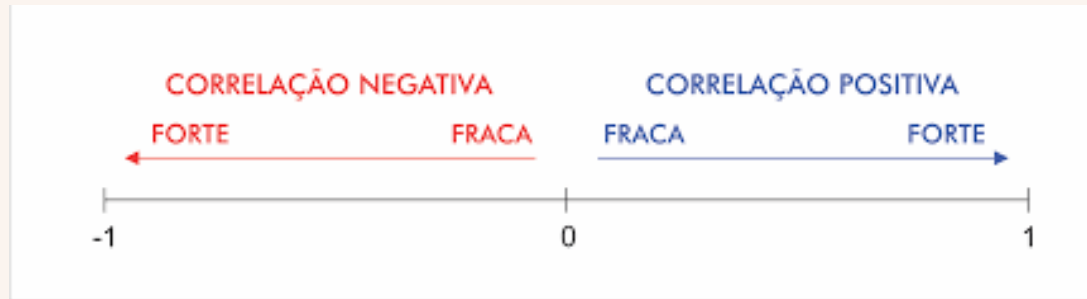
Como é que X produz um efeito sobre Y?

Exemplo: educação e participação política.

CAUSALIDADE

- As pessoas com mais escolaridade tendem a participar mais.
- Para cada ano de estudo há um incremento na capacidade de obter e processar informações políticas, bem como nas formas de encaminhar demandas. Desta forma, providos de mais informação sobre a realidade política e conhecendo melhor os canais adequados de fazer valer a sua vontade, os cidadãos com maiores níveis educacionais possuem uma gama maior de repertórios disponíveis para manifestar a sua vontade.

PARÂMETROS



Pressupostos:

1. As variáveis devem ser contínuas.
2. Devem possuir uma distribuição normal.
3. Independência entre as observações.

O coeficiente de correlação de Pearson não diferencia X e Y de Y e X.

Alterar a mensuração das variáveis não altera o valor da correlação. Se eu testei a relação entre quilos e decilitros não encontrarei diferenças entre toneladas e litros.

Os *outliers* afetam o valor do coeficiente!

CORRELAÇÃO DE PEARSON

CORRELAÇÃO NO R

Vamos agora abrir a base sen2018.csv

```
##Carregar a Base de Dados##  
sen2018 <- read.csv2("sen2018.csv",  
                     encoding = "latin1")
```

Problema: maior arrecadação implicou em mais votos ao senado federal nas eleições de 2018?

X = total de arrecadação.

Y = votos

Correlação no R: Existem duas funções que computam o coeficiente de correlação. O primeiro é o comando `cor()`, que retorna apenas o coeficiente. O segundo é o `cor.test()` que retorna tanto o coeficiente quanto seu nível de significância.

```
> cor(sen2018$TOTAL_RECEITAS, sen2018$votos)  
[1] 0.4295534
```

CORRELAÇÃO NO R

```
> cor.test(sen2018$TOTAL_RECEITAS, sen2018$votos)
```

Pearson's product-moment correlation

data: sen2018\$TOTAL_RECEITAS and sen2018\$votos

t = 8.3616, df = 309, p-value = 2.144e-15

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.3343072 0.5161140

sample estimates:

cor

0.4295534

CORRELAÇÃO NO R



No exemplo anterior encontramos uma correlação moderada de positiva ($r=0,43$) entre a arrecadação e o total de votos dos candidatos ao senado. Mas será que este valor é adequado?



Para melhorar a interpretação do nosso resultado, vamos plotar um gráfico de dispersão para identificar possíveis *outliers*.

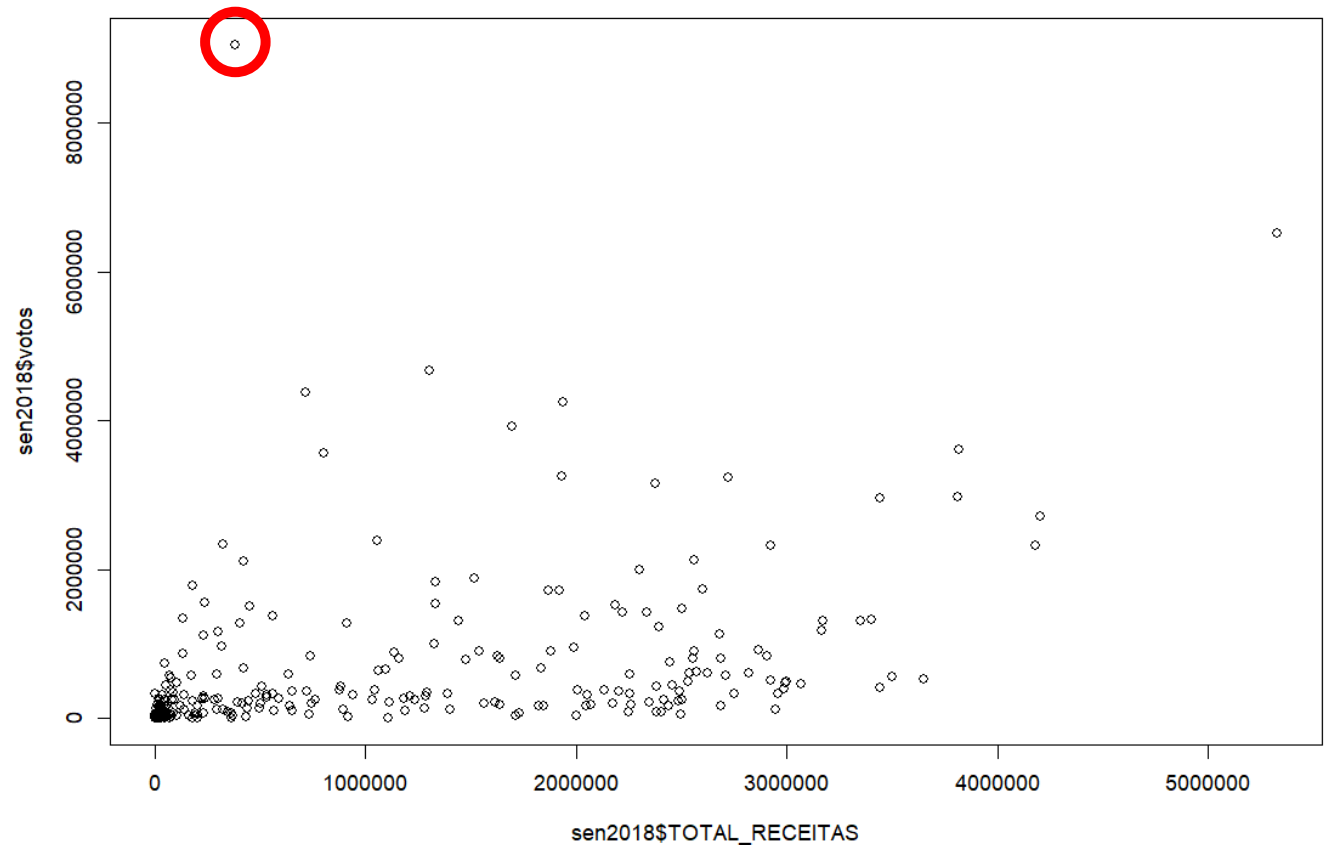
CORRELAÇÃO NO R

Aparentemente temos um outlier. Ele foi o que mais recebeu votos, mas arrecadou pouco dinheiro.

Vamos checar qual caso é este.

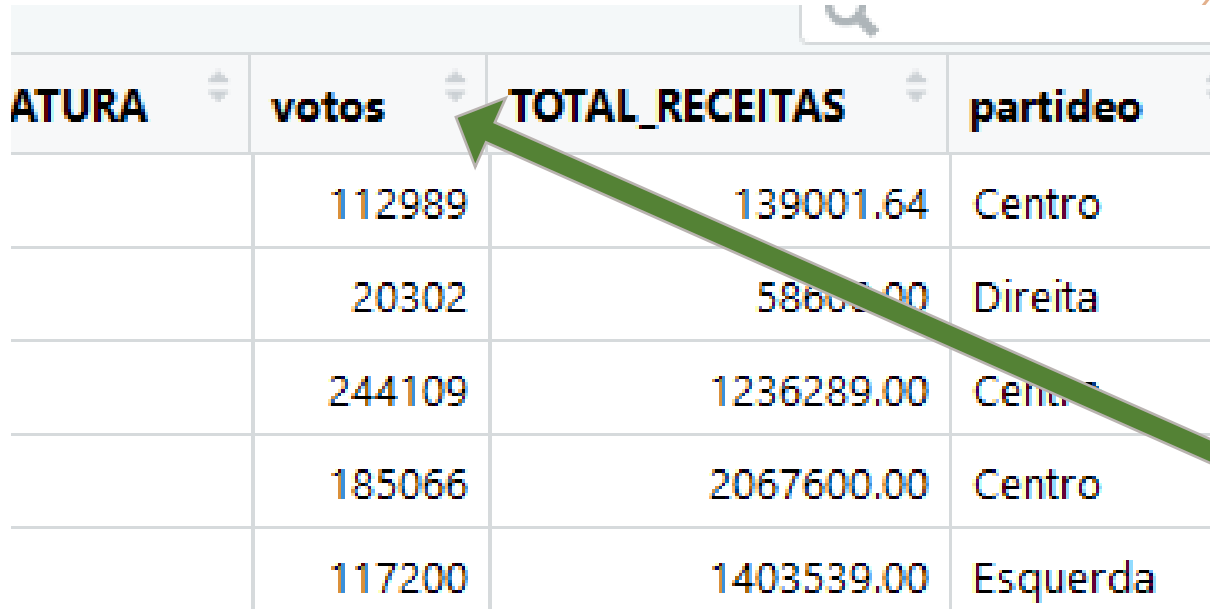
Vamos visualizar o banco de dados com o comando

```
View(sen2018)
```



CORRELAÇÃO NO R

- Achem a coluna “votos” e cliquem na seta para baixo.



| ATURA | votos | TOTAL_RECEITAS | partideo |
|-------|--------|----------------|----------|
| | 112989 | 139001.64 | Centro |
| | 20302 | 58600.00 | Direita |
| | 244109 | 1236289.00 | Centro |
| | 185066 | 2067600.00 | Centro |
| | 117200 | 1403539.00 | Esquerda |

Os votos passaram a ordenar os casos do menor para o maior. Rolem até o final e identifiquem quem é o nosso outlier...

CORRELAÇÃO NO R

O Major Olímpio (PSL – SP) fez 9.039.717 votos, mas declarou ter arrecadado apenas 380.323,96 reais.

Agora vamos identificar qual o número do caso ele representa.

Arrastem a barra de rolagem para a esquerda, a primeira coluna sem identificação. Ela apresenta o valor que o R atribui a cada linha.

Vamos excluir este caso (285) e recalcular o coeficiente de correlação.

```
sen2018 ← sen2018[-285, ]
```

CORRELAÇÃO NO R

Pearson's product-moment correlation

```
data: sen2018$TOTAL_RECEITAS and sen2018$votos
t = 10.317, df = 308, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4190319 0.5851373
sample estimates:
      cor
0.5067732
```

Nossa base de dados possui 311 casos. Ao excluir um único outlier, o valor de r foi de 0,43 para 0,51!

Mas ainda podemos melhorar nossa análise. Em geral, quando trabalhamos com dinheiro, o seu efeito tende a ser não linear, ou seja, a partir de um ponto ele passa a não produzir o mesmo efeito.

Quando isso ocorre, usamos uma função logarítmica para reduzir este efeito.

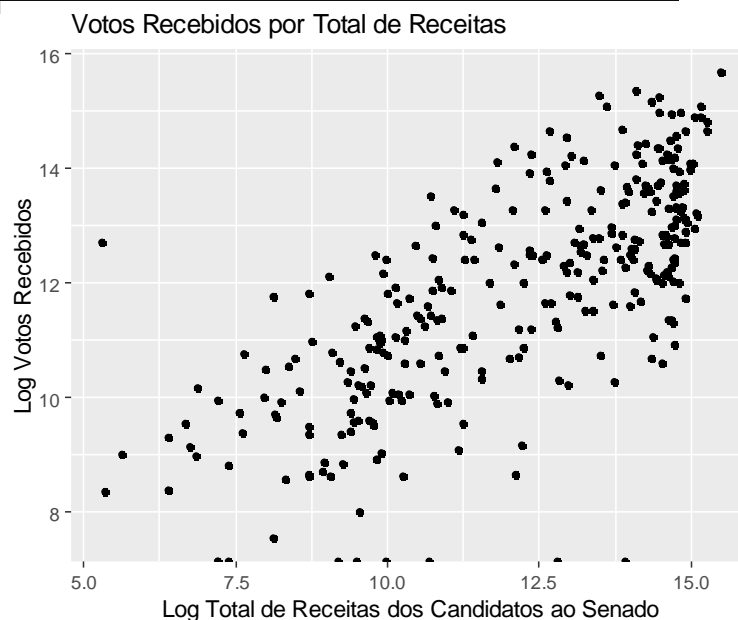
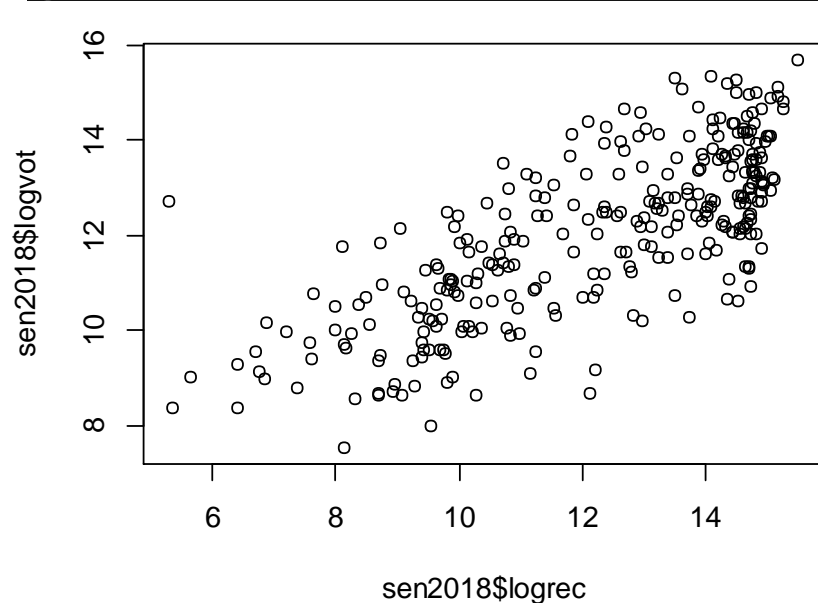
No caso da correlação, não estamos mais trabalhando com os efeitos reais das variáveis, mas sim com a função logarítmica. Isso deve ficar claro na hora de reportar os dados.

. VER NO R

LOGARITMAR

"Logaritmar" é o processo de calcular o logaritmo de um número ou variável, que é a operação inversa da exponenciação. O logaritmo é útil para transformar distribuições, linearizar relações matemáticas, resolver equações complexas e comparar números em escalas menores, tendo aplicações em várias áreas, incluindo estatística, modelagem matemática, ciência e economia.

```
###LOG###  
sen2018$logrec ← log(sen2018$TOTAL_RECEITAS)  
sen2018$logvot ← log(sen2018$votos)  
plot(sen2018$logrec, sen2018$logvot)
```



ok

CORRELAÇÃO NO R



Uma Matriz de Correlação correlaciona todas as variáveis e as exibe em uma única tabela.



Para isto devemos separar as variáveis que queremos correlacionar em uma base de dados separada.



Para isto vamos usar o pacote Hmisc

```
install.packages("Hmisc")  
library(Hmisc)
```



Primeiro vamos selecionar em uma base separada apenas as variáveis que desejamos correlacionar.

CORRELAÇÃO NO R

```
###separar a base###
sen2018cor <- sen2018[c("IDADE_DATA_POSSE",
                        "logvot", "logrec")]

###retirar os 0's###
sen2018cor <- subset(sen2018cor, logvot > 0)

###matriz de correlação###
install.packages("Hmisc")
library(Hmisc)
mcor <- rcorr(as.matrix(sen2018cor))
mcor
```

output

```
> mcor <- rcorr(as.matrix(sen2018cor))
> mcor
```

| | IDADE_DATA_POSSE | logvot | logrec |
|------------------|------------------|--------|--------|
| IDADE_DATA_POSSE | 1.00 | 0.20 | 0.32 |
| logvot | 0.20 | 1.00 | 0.75 |
| logrec | 0.32 | 0.75 | 1.00 |

n= 301

P

| | IDADE_DATA_POSSE | logvot | logrec |
|------------------|------------------|--------|--------|
| IDADE_DATA_POSSE | | 0.0005 | 0.0000 |
| logvot | 0.0005 | | 0.0000 |
| logrec | 0.0000 | 0.0000 | |

CORRELAÇÃO NO R

Podemos criar um gráfico de calor para exibir a matriz de correlação. Vamos usar o pacote `corrplot()`

```
install.packages("corrplot")
```

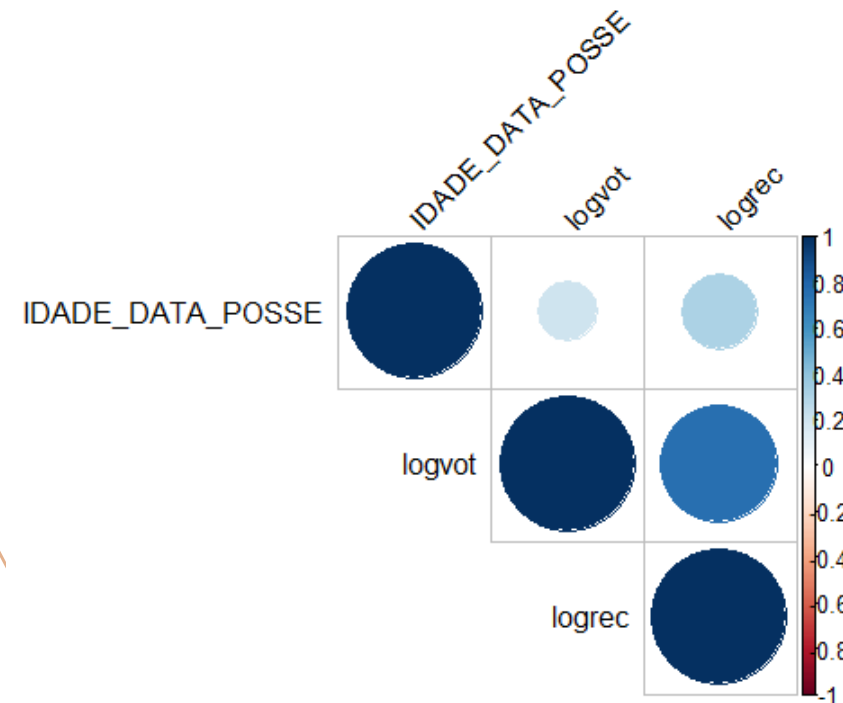
```
library(corrplot)
```

```
corrplot(mcor$r, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
```

Type = "upper" exibe apenas a metade de cima da matriz, já que a outra parte são seus valores espelhados.

Order = "hclust" ordena os coeficientes do maior para o menor

tl.col e tl.str indicam a cor e o tamanho das rotações.



RHO DE SPEARMAN

Como o cálculo do Rho de Spearman é feito a partir do ordenamento (rank) das frequências, ele pode ser empregado com variáveis contínuas.

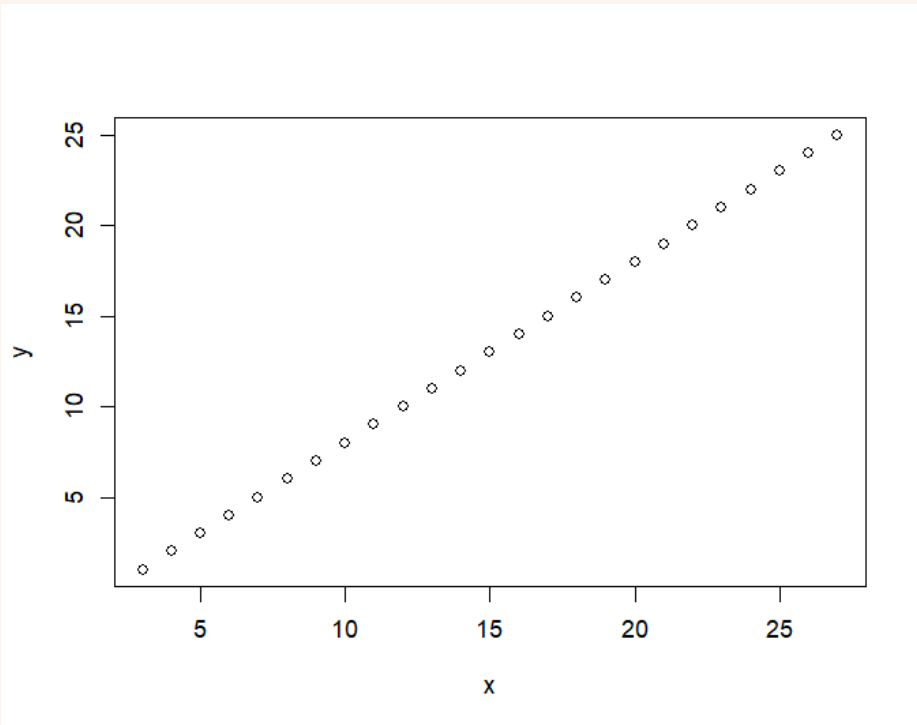
Ele é menos rigoroso que outros testes, logo há uma tendência de apresentar coeficientes mais altos

Interpretando os resultados.

- A primeira coisa que devemos observar é o teste de hipótese expresso pelo valor de p (p-value).
- Ele teste se a hipótese nula (o valor de rho verdadeiro é igual a zero) deve ser mantida ou rejeitada.
- Assumimos que um valor de $p < 0,05$ é aceitável.
- No caso do teste anterior, $p < 0,001$, logo devemos rejeitar a hipótese nula.

RHO DE SPEARMAN COMPARADO COM PEARSON

Em seguida deve observar o sinal do



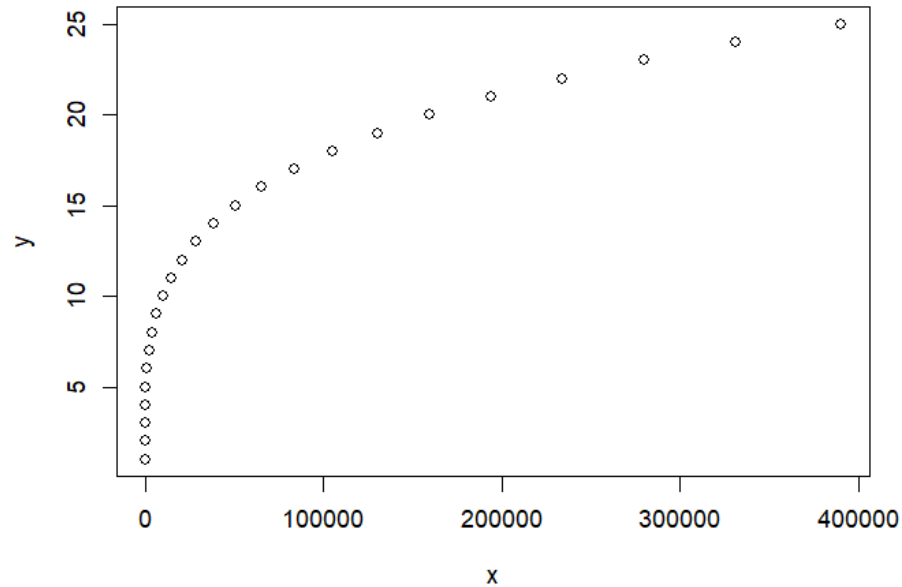
Pearson's product-moment correlation

```
data: x and y
t = Inf, df = 23, p-value < 0.000000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 1 1
sample estimates:
cor
 1
```

Spearman's rank correlation rho

```
data: x and y
S = 0, p-value = 0.0000003196
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
 1
```


RHO DE SPEARMAN COMPARADO COM PEARSON



Pearson's product-moment correlation

```
data: x and y
t = 8.522, df = 23, p-value = 0.00000001426
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7265273 0.9421742
sample estimates:
      cor
0.8714796
```

Spearman's rank correlation rho

```
data: x and y
S = 0, p-value = 0.00000003196
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
1
```

RHO DE SPEARMAN

No primeiro exemplo temos uma relação linear perfeita entre X e Y. Neste caso, os valores dos coeficientes de correlação de Pearson e Spearman coincidiram.

Já no segundo exemplo (onde $X = Y^4$), podemos observar uma redução no valor do coeficiente de Pearson, mas ainda se manteve a correlação perfeita em Spearman.

A diferença ocorre porque na correlação de Pearson usamos os valores das variáveis para o cálculo. Já na correlação de Spearman não são empregados os valores, mas o ranking em que eles estão dispostos.

TAU DE KENDALL

A correlação de Kendall também é calculada a partir do ranking das categorias.

Mas diferente da correlação de Spearman, o cálculo da correlação de Kendall leva em consideração os pares discordantes.

Por levar em consideração os pares discordantes no cálculo, a correlação de Kendall é mais robusta que a correlação de Spearman.

Isto faz com que o valor de Tau seja menor que o Valor de Rho

Assim, é preferível usar a correlação de Kendall sempre que possível.



TAU DE KENDALL

No R, usamos a mesma função `cor.test()`, mas no método especificamos "kendall".

```
> ##Correlação de Kendall##  
> cor.test(bd$avalmil, bd$confmil, method = "kendall", na.rm = T)
```

Kendall's rank correlation tau

```
data: bd$avalmil and bd$confmil  
z = 26.64, p-value < 0.000000000000000022  
alternative hypothesis: true tau is not equal to 0  
sample estimates:  
      tau  
0.4653411
```

```
> ##Correlação de Spearman##  
> cor.test(bd$avalmil, bd$confmil, method = "spearman", na.rm = T)
```

Spearman's rank correlation rho

```
data: bd$avalmil and bd$confmil  
s = 959274469, p-value < 0.000000000000000022  
alternative hypothesis: true rho is not equal to 0  
sample estimates:  
      rho  
0.5318481
```

Como vimos anteriormente, a correlação de Kendall é mais robusta que a correlação de Spearman.

Desta forma, seus valores tendem a ser ligeiramente menores.

Se possível, entre Kendall e Spearman, Kendall é preferível.

A series of thin, light brown lines of varying lengths and orientations intersecting to form a complex, abstract geometric pattern on the left side of the slide.

OBRIGADA ! =)