CIÊNCIAS AMBIENTAIS / BIOLÓGICAS / DA NATUREZA

(BIO)ESTATÍSTICA

Profª. Letícia Raposo
profleticiaraposo@gmail.com

TESTANDO ASSOCIAÇÕES ENTRE VARIÁVEIS CATEGÓRICAS

OBJETIVOS

- Apreciar a base conceitual do teste qui-quadrado;
- Ser capaz de escolher entre uma análise direta pelo qui-quadrado e pelo teste exato de Fisher;
- Saber como conduzir uma análise relevante usando o R e como interpretar os resultados;
- Ser capaz de elaborar tabelas de contingência e escrever os resultados;
- Entender o uso das análises das tabelas de contingência na literatura publicada.

INTRODUÇÃO

A associação entre duas variáveis categóricas pode ser resumida em uma tabela de contingência.

- As categorias de uma variável estão listadas na primeira linha da tabela e as categorias da segunda variável na primeira coluna.

Ex: Podemos lançar a hipótese de que motoristas do sexo masculino têm uma probabilidade maior de sofrer um acidente de carro do que motoristas do sexo feminino.

	Sem acidente	Acidente	Total
Mulheres	51	5 (0,09)	56
Homens	29	15 (0,34)	44
Total	80	20	100

INTRODUÇÃO

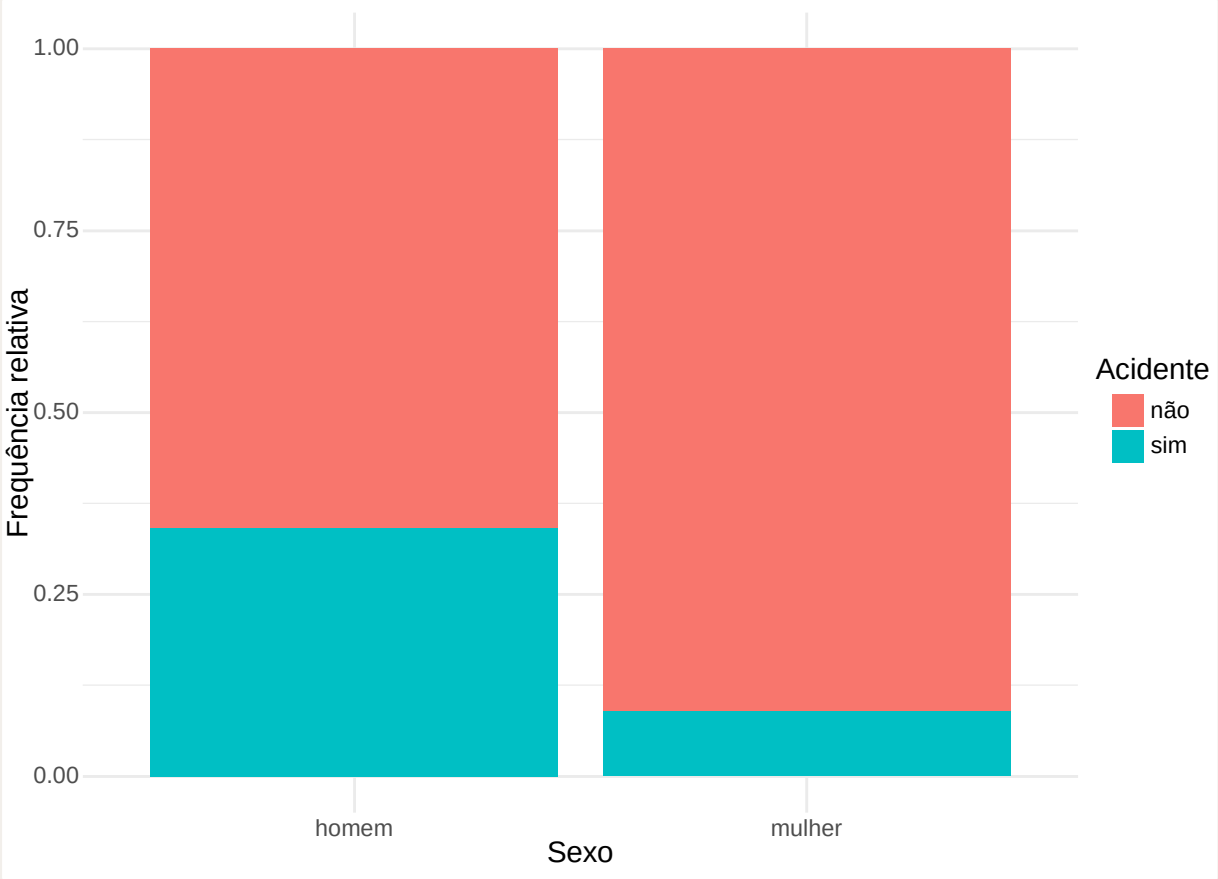
- Nós temos somente uma amostra, portanto as diferenças aparentes poderiam ser produto do erro amostral.
- Estatística inferencial: testar a probabilidade de que essa associação possa ter ocorrido por acaso → estatística do qui-quadrado χ^2.
- Calculamos a probabilidade de os dados observados serem gerados no caso de a hipótese nula ser verdadeira (não há associação entre as variáveis).
- Quanto maior a estatística χ^2, menor a probabilidade da hipótese nula ser verdadeira. Se a probabilidade for menor que 0,05, então rejeitamos H0 e aceitamos que existe associação na população.

A LÓGICA DA ANÁLISE DAS TABELAS DE CONTINGÊNCIA

- A estatística χ^2 é baseada na diferença entre as frequências observadas e as esperadas.
- Para cada célula, a frequência esperada é subtraída da observada. O número resultante é, então, elevado ao quadrado (devido a presença de valores negativos). A seguir, dividimos o resultado pela frequência esperada. Então, simplesmente somamos os números calculados em cada célula, e isso nos dá a estatística χ^2.

	Sem acidente	Acidente
Mulheres	$(51-44,8)^2/44,8=0,86$	$(5-11,2)^2/11,2=3,43$
Homens	$(29-35,2)^2/35,2=1,09$	$(15-8,8)^2/8,8=4,37$
$\chi^2=0,86+3,43+1,09+4,37=9,75$		

A LÓGICA DA ANÁLISE DAS TABELAS DE CONTINGÊNCIA



EXECUTANDO A ANÁLISE NO R

```
> qui <- chisq.test(acidentes$sexo, acidentes$acidente)
> qui

        Pearson's Chi-squared test with Yates' continuity correction

data:  acidentes$sexo and acidentes$acidente
X-squared = 8.2412, df = 1, p-value = 0.004095

> qui$observed
              acidentes$acidente
acidentes$sexo não sim
        homem   29  15
        mulher  51   5
> qui$expected
              acidentes$acidente
acidentes$sexo  não  sim
        homem  35.2  8.8
        mulher 44.8 11.2
```

(Número de linhas – 1) X (Número de colunas – 1)

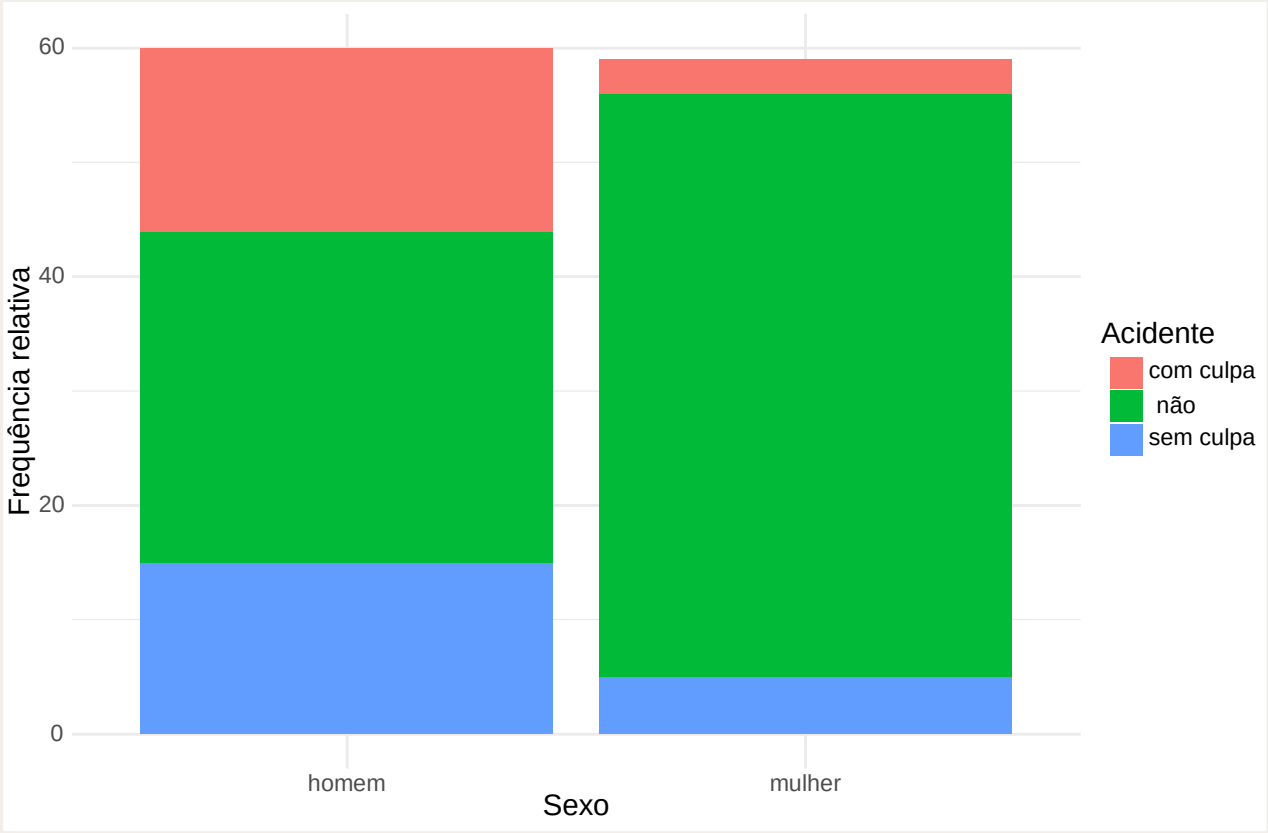
Existe uma associação significativa entre o gênero e o envolvimento em acidentes de carro (χ^2 (gl = 1, n = 100) = 8,24, p = 0,004). Os homens têm uma probabilidade significantemente maior de ter se envolvido em um acidente de carro (34,1%) do que as mulheres (8,9%).

GRANDES TABELAS DE CONTINGÊNCIA

- Podemos estender a análise para variáveis com mais categorias (p. ex., uma tabela 3 x 2 ou 4 x 3 ou o que você quiser *versus* o que você quiser).
- Vamos imaginar agora a variável do histórico de acidente com três categorias: nenhum acidente, acidentes sem culpa, acidentes com culpa.

	Sem acidente	Acidente sem culpa	Acidente com culpa	Total
Mulheres	51	5	3	59
Homens	29	15	16	60
Total	80	20	19	119

GRANDES TABELAS DE CONTINGÊNCIA



GRANDES TABELAS DE CONTINGÊNCIA

```
> qui <- chisq.test(acidentes_culpa$sexo, acidentes_culpa$acidente)
> qui

        Pearson's Chi-squared test

data:  acidentes_culpa$sexo and acidentes_culpa$acidente
X-squared = 19.938, df = 2, p-value = 4.684e-05

> qui$observed
                    acidentes_culpa$acidente
acidentes_culpa$sexo com culpa não sem culpa
              homem         16  29        15
              mulher         3  51         5
> qui$expected
                    acidentes_culpa$acidente
acidentes_culpa$sexo com culpa      não sem culpa
              homem   9.579832 40.33613 10.084034
              mulher  9.420168 39.66387  9.915966
```

O resultado significativo ainda implica que o histórico de acidentes e o gênero estão associados, mas, na interpretação de uma tabela grande, isso já não fica mais tão claro.

A significância do χ^2 não nos diz se a distribuição de homens e mulheres difere entre todas as categorias da variável acidentes ou somente em algumas delas.

GRANDES TABELAS DE CONTINGÊNCIA

Uma forma de resolver o problema é agregar as categorias até se obter uma tabela 2 x 2.

- Por exemplo, você pode agregar as categorias culpa e sem culpa em uma única categoria de acidentes caso todos os acidentes sejam de nosso interesse teórico, independentemente da culpa.
- Se você deseja incluir mais variáveis, a regressão logística pode ser útil para algumas análises desse tipo e, também, possui a vantagem de que tanto variáveis contínuas quanto categóricas podem ser incluídas como preditoras.

SUPOSIÇÕES DA ANÁLISE DE TABELAS DE CONTINGÊNCIA

- As categorias para cada variável devem ser mutuamente exclusivas.
  - Cada participante pode ser colocado apenas em uma categoria de cada variável → não é adequada para delineamentos intra-participantes.
- Existe pelo menos uma observação em cada célula da tabela.
  - Se isso for um problema, então você pode resolvê-lo agregando as categorias, caso aquelas que forem agregadas formem categorias significativas.
- As frequências esperadas não estão abaixo de cinco em mais de 20% das células da tabela de contingência.
  - Felizmente, existe uma estatística teste alternativa que não é vulnerável a frequências esperadas baixas: o teste exato de Fisher.

TESTE EXATO DE FISHER

- Quando pelo menos um dos valores esperados for menor que 5, recomenda-se o uso do Teste Exato de Fisher.
- O Teste Exato de Fisher é recomendado para amostras pequenas ($N < 20$).

TESTE EXATO DE FISHER

Uma mulher britânica afirmou ser capaz de distinguir se leite ou chá foi adicionado à xícara primeiro. Para testar, ela recebeu 8 xícaras de chá, das quais quatro receberam primeiro o leite. A hipótese nula é a de que não há associação entre a verdadeira ordem da adição e o palpite da mulher, a alternativa de que existe uma associação.

```
> cha <-
+   matrix(c(3, 1, 1, 3),
+          nrow = 2,
+          dimnames = list(Guess = c("Leite", "Chá"),
+                          Truth = c("Leite", "Chá")))
> chisq.test(cha) # veja e mensagem de erro

        Pearson's Chi-squared test with Yates' continuity correction

data:  cha
X-squared = 0.5, df = 1, p-value = 0.4795

Warning message:
In chisq.test(cha) : Aproximação do qui-quadrado pode estar incorreta
```

TESTE EXATO DE FISHER

Uma mulher britânica afirmou ser capaz de distinguir se leite ou chá foi adicionado à xícara primeiro. Para testar, ela recebeu 8 xícaras de chá, das quais quatro receberam primeiro o leite. A hipótese nula é a de que não há associação entre a verdadeira ordem da adição e o palpite da mulher, a alternativa de que existe uma associação.

```
> fisher.test(cha)

        Fisher's Exact Test for Count Data

data:  cha
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.2117329 621.9337505
sample estimates:
odds ratio
  6.408309
```

AVALIANDO A CONCORDÂNCIA: TÉCNICAS CORRELACIONAIS

OBJETIVOS

- Obter um entendimento conceitual da análise correlacional;
- Ser capaz de decidir quando usar um teste paramétrico (r de Pearson) e quando usar o não paramétrico equivalente (rô de Spearman);
- Entender as situações em que você pode sugerir causalidade ao usar a análise de correlação;
- Aprender como executar a análise correlacional bivariada;
- Aprender a interpretar os resultados dos pesquisadores que usaram a análise correlacional em suas publicações.

INTRODUÇÃO

- As técnicas correlacionais observam os relacionamentos ou as associações entre as variáveis, e não olham para as diferenças entre médias.
- A análise correlacional é ideal quando os pesquisadores estão observando um comportamento que ocorre naturalmente → eles não alocam pessoas aos grupos.

RELACIONAMENTOS BIVARIADOS

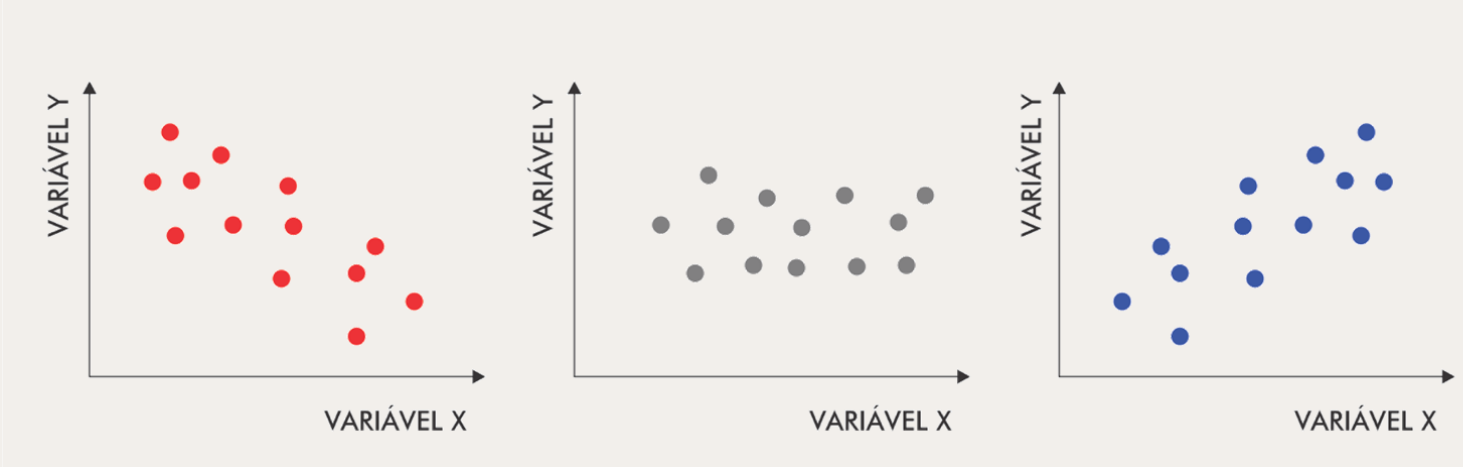
- Um relacionamento bivariado ocorre quando uma variável mostra uma associação ou correlação com uma segunda variável.
- Técnicas de correlação bivariada avaliam a força e a magnitude da associação (relacionamento) entre duas variáveis, e o valor-p associado nos mostra se esse relacionamento ocorre devido ao erro amostral (ou acaso).
- Técnicas correlacionais não são usadas para avaliar diferenças entre variáveis.

RELACIONAMENTOS BIVARIADOS

- As hipóteses são geralmente direcionais.
- Quando os pesquisadores formulam uma hipótese direcional, pode-se utilizar um nível de significância unilateral na avaliação dos resultados. Se eles simplesmente preveem um relacionamento, mas não têm nenhuma razão lógica para prever sua direção, então será utilizado um nível de significância bilateral.

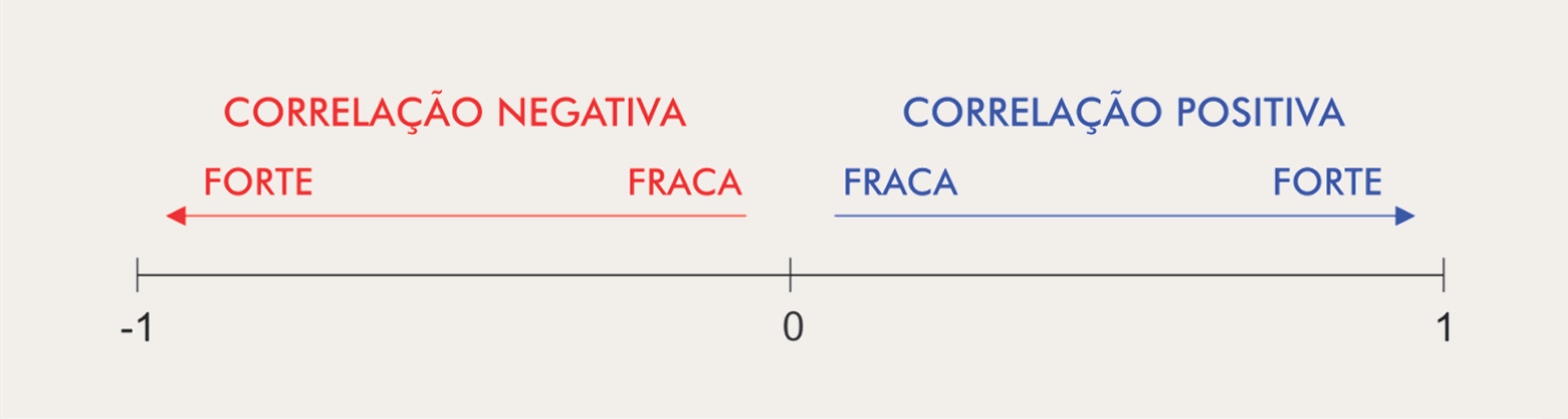
RELACIONAMENTOS BIVARIADOS

- Quando os pesquisadores executam uma análise correlacional, eles observam os diagramas de dispersão para ter uma ideia geral dos relacionamentos.
- A força do relacionamento entre as variáveis é avaliada não pelo diagrama de dispersão, mas por um teste estatístico.
  - Teste paramétrico: correlação produto momento de Pearson;
  - Teste não paramétrico: rô de Spearman.



COEFICIENTE DE CORRELAÇÃO DE PEARSON

- O coeficiente de correlação de Pearson (r), também chamado de correlação linear ou r de Pearson, é um grau de relação entre duas variáveis quantitativas e exprime o grau de correlação através de valores situados entre -1 e 1.
- Cohen dividiu em valores fracos (0,1 a 0,3), moderados (0,4 a 0,6) ou fortes (0,7 a 0,9).



EXEMPLO

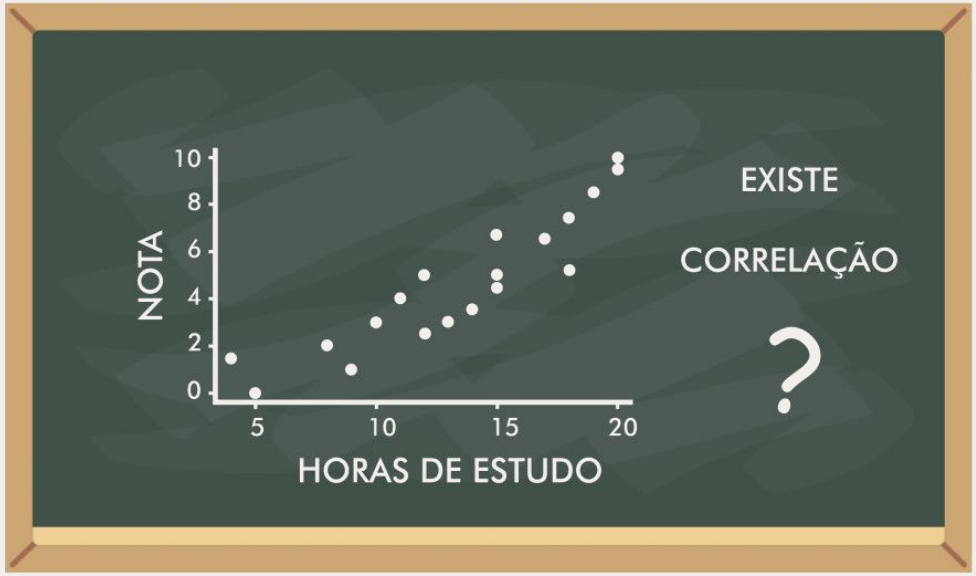
- Um professor percebe que alguns de seus alunos não estão apresentando bom desempenho nas provas e, percebendo que estes não estavam dedicando tempo suficiente aos estudos, decide fazer uma pequena experiência com a turma.
- Ele então pede que os alunos informem o tempo que cada um dedicou em casa ao estudo do conteúdo cobrado e monta a tabela ao lado.

http://www.abgconsultoria.com.br/blog/coeficientes-de-correlacao/

Aluno	Horas	Nota
1	20	9,5
2	12	2,5
3	14	3,6
4	15	6,7
5	18	5,2
6	9	1
7	5	0
8	4	1,5
9	8	2
10	13	3
11	14	3,5
12	15	4,5
13	19	8,5
14	18	7,5
15	12	5
16	11	4
17	10	3
18	15	5
19	17	6,5
20	20	10

EXEMPLO

Ele explica aos alunos que se existe uma relação entre as horas de estudo com as notas da prova, isso poderia facilmente ser observado em um gráfico. Utilizando o eixo X para as horas de estudo e o eixo Y para a nota na prova, marca no gráfico a nota e o tempo de estudo de cada aluno.

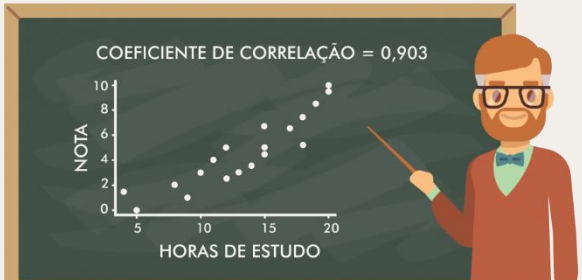


EXEMPLO

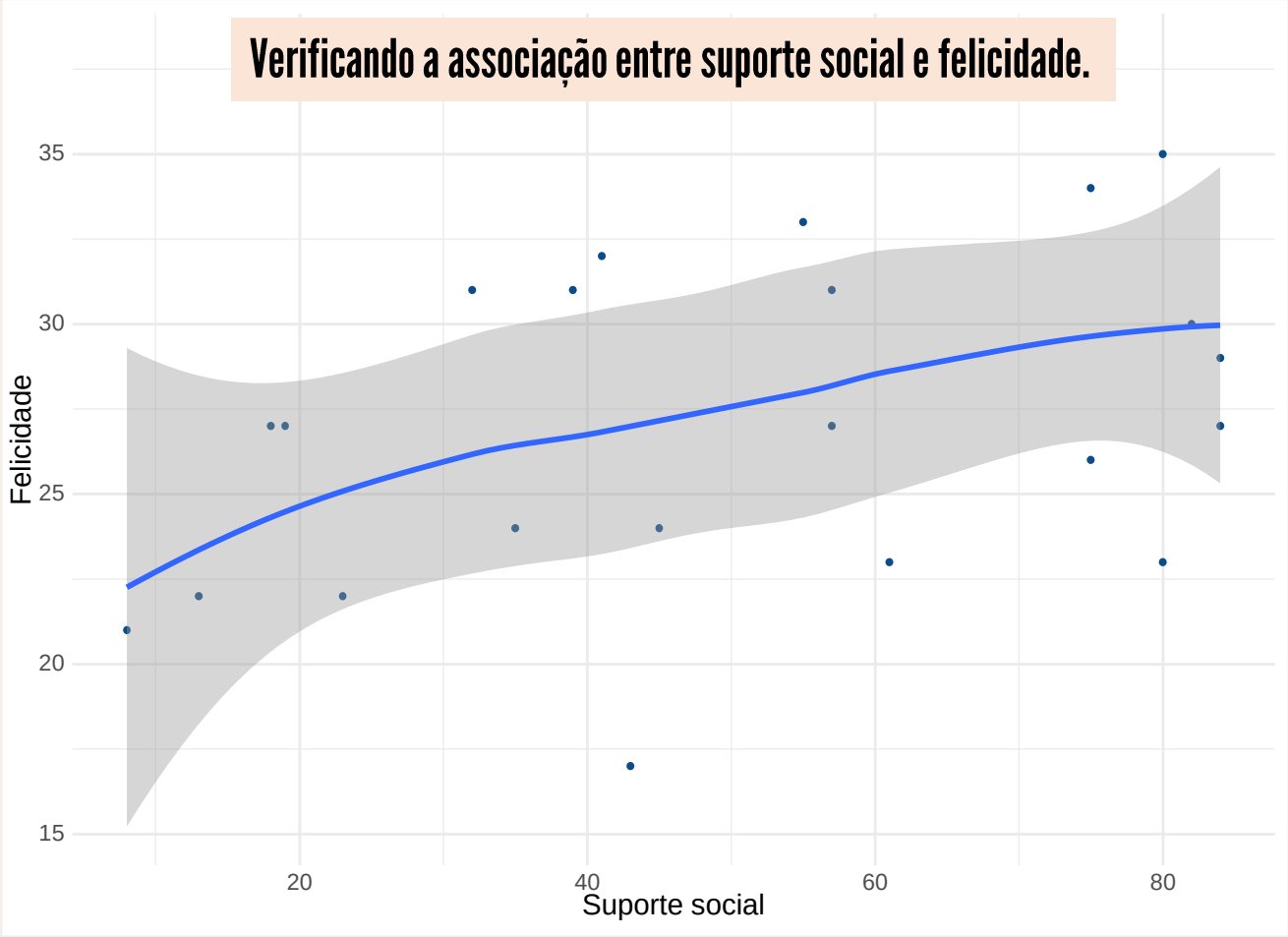
- Visualmente parece haver alguma relação do tempo de estudo e a nota da prova, pois quanto maior o tempo de estudo, maior tende a ser a nota do aluno. Mas como confirmar e quantificar essa relação?
- Além de calcular o coeficiente de correlação, existem testes estatísticos que permitem avaliar as hipóteses de que o coeficiente é igual a zero (hipótese nula) e de que ele é diferente de zero (hipótese alternativa).
- O professor então decide utilizar o coeficiente de correlação de Pearson e chegou ao valor de r = 0,903, com p-valor de 0,000.

EXEMPLO

- Se o coeficiente de correlação de Pearson sustenta resultados entre -1 e 1, podemos dizer que nesse caso existe uma relação positiva entre horas de estudo e a nota da prova, como era de se esperar.
- Além disso, o p-valor do teste de Pearson de 0,000 indica a rejeição da hipótese de que o coeficiente de correlação seja igual a zero, indicando que existe uma relação significativa entre as variáveis testadas.
- Após apresentar evidências de que quanto mais um aluno estude em casa, maior tende a ser sua nota na prova, o professor espera que os alunos se dediquem mais aos estudos!



COEFICIENTE DE PEARSON NO R



COEFICIENTE DE PEARSON NO R

Verificando a associação entre suporte social e felicidade.

```
> shapiro.test(suporte$socialsupport)

        Shapiro-Wilk normality test

data:  suporte$socialsupport
W = 0.92205, p-value = 0.07373

> shapiro.test(suporte$happiness)

        Shapiro-Wilk normality test

data:  suporte$happiness
W = 0.98354, p-value = 0.9567
```

COEFICIENTE DE PEARSON NO R

```
> cor.test(suporte$happiness, suporte$socialsupport)

        Pearson's product-moment correlation

data:  suporte$happiness and suporte$socialsupport
t = 2.2454, df = 21, p-value = 0.03564
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.03396012 0.72137010
sample estimates:
      cor
0.4400029
```

O suporte social e a felicidade mostram um relacionamento positivo (moderado) estatisticamente significativo (r = 0,440, p = 0,018).

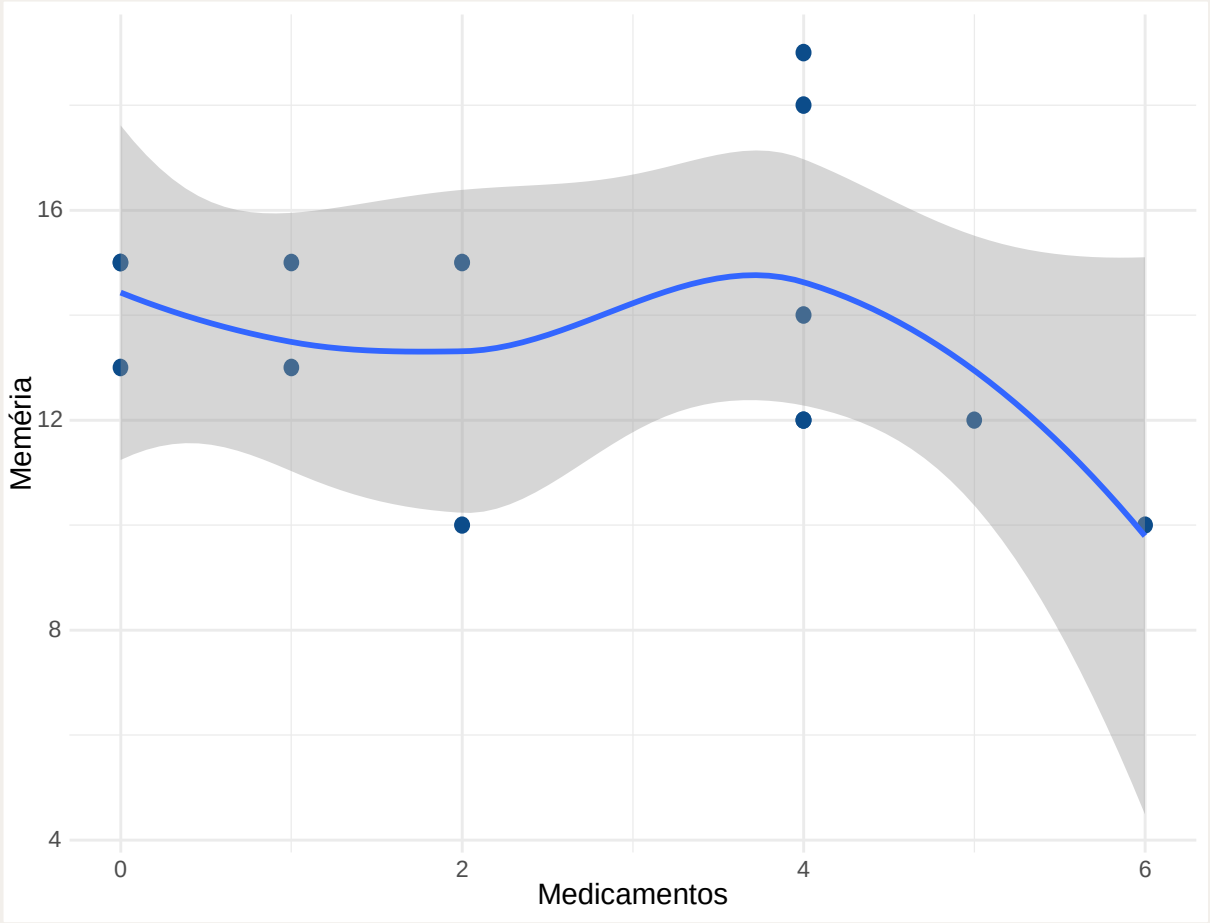
COEFICIENTE DE CORRELAÇÃO DE SPEARMAN

- Denominado pela letra grega rho (ρ), o coeficiente de correlação de postos de Spearman é uma medida de correlação não paramétrica também avaliado no intervalo entre -1 e 1.
- Ao contrário do coeficiente de Pearson, o coeficiente de Spearman não exige a suposição de que a relação entre as variáveis seja linear, nem requer que as mesmas sejam quantitativas – pode inclusive ser utilizado para verificar relação entre variáveis medidas no nível ordinal.

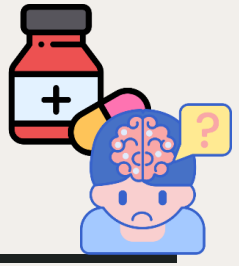
COEFICIENTE DE CORRELAÇÃO DE SPEARMAN NO R

- Um pesquisador pegou informações sobre quantas medicações diferentes 14 pacientes tomaram na semana anterior e as correlacionou com a medida de memória.
- De acordo com a literatura, o autor espera que quanto maior o número de medicamentos, pior será a memória do indivíduo, e, então, ele opta por usar um teste unilateral.

COEFICIENTE DE CORRELAÇÃO DE SPEARMAN NO R



COEFICIENTE DE CORRELAÇÃO DE SPEARMAN NO R



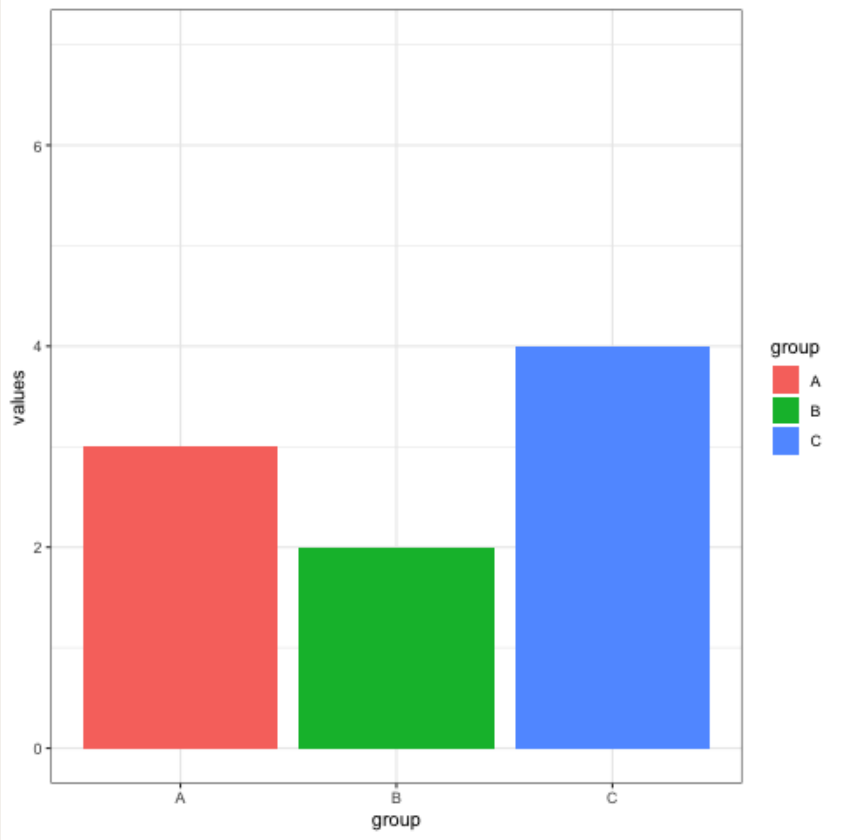
```
> cor.test(memoria$medications, memoria$memory, method = "spearman")

        Spearman's rank correlation rho

data:  memoria$medications and memoria$memory
S = 612.27, p-value = 0.2261
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.3456451
```

O coeficiente de correlação é negativo (como previsto), mas fraco e estatisticamente não significativo em qualquer nível aceitável de significância. Portanto, o pesquisador conclui que o número de medicamentos não estava significativamente relacionado à memória fraca.

ARTE DO DIA FEITA EM R



https://www.r-graph-gallery.com/288-animated-barplot-transition.html

REFERÊNCIAS BIBLIOGRÁFICAS

- BARBETTA, Pedro Alberto. Estatística aplicada às ciências sociais. Ed. UFSC, 2008.
- DANCEY, Christine P.; REIDY, John G.; ROWE, Richard. Estatística Sem Matemática para as Ciências da Saúde. Penso Editora, 2017.
- MAGNUSSON, Willian E. Estatística [sem] matemática: a ligação entre as questões e a análise. Planta, 2003.