Estatística Instrumental: Análise Bivariada

Alexandre Costa & Leo Am

Estatística Instrumental: Análise Bivariada

Alexandre Costa & Leo Am

Esse livro está à venda em http://leanpub.com/bivariada

Essa versão foi publicada em 2022-03-26

Conteúdo

1. Análise bivariada

A análise estatística pode buscar avaliar a relação entre variáveis e se perguntar se há potencial interferência de um fenômeno em outro. Quando buscamos avaliar a relação entre duas variáveis da mesma população ou amostra, podemos qualificar essa análise bivariada. Esse tipo de abordagem pode responder a questões como:

- A distribuição para determinado relator (M1) aumenta as chances de um processo ser julgado procedente?
- Juízes eleitos atuam de forma a privilegiar seus potenciais eleitores, em detrimento de atores que não interferem em sua eleição?
- Juízas mulheres julgam de forma diferente dos homens casos de agressão sexual?

Tais questões levam a análise bivariadas porque elas exigem a comparação entre duas variáveis:

- relatoria de M1 e índices de procedência;
- juiz ser da classe eleito e valor de indenização de certos processos ou
- gênero do juiz e resultados processuais.

Em um primeiro momento, você deve avaliar se existe uma forte correlação entre esses fenômenos, ou seja, se eles variam de maneira convergente em uma determinada população. Este estudo muitas vezes dependerá da definição de amostras adequadas, que você já sabe fazer. No presente texto você aprenderá a fazer correlações simples e a compreender análises feitas em textos jurídicos.

Uma vez que você localize uma forte correlação, você pode dar um passo adiante e perguntar se existe uma relação causal entre eles. Essa segunda etapa da análise é complexa e ultrapassará os limites desta abordagem introdutória e instrumental, pois nosso objetivo é que você compreenda estudos que usem essas ferramentas

(aprender a realizar de forma autônoma essas análises está além dos limites da abordagem viável no presente curso).

Nete curso, nós nos focaremos em entender um pouco da análise de correlação, que pode por si só ajudar a refutar ligações entre fenômenos, ou confirmar o interesse de uma linha de pesquisa a ser aprofundada.

1.1 Variáveis dependentes e independentes

Na avaliação estatística da causalidade, convencionou-se descrever essas análises bivariadas como formas de investigar se a variação de uma característica pode ser explicada pela variação de outra medida. Por exemplo, podemos investigar se:

1. Se a mortalidade reduzida de uma doença pode ser explicada pela aplicação de uma vacina;
2. Se o tempo de cura de uma doença pode ser explicada pela utilização de certo fármaco;
3. Se a duração de certos processos pode ser explicada pelas características de seus autores.

A medida que se deseja explicar (tempo de cura, mortalidade, duração processual) é chamada de variável dependente, no sentido de que o que se investiga é exatamente se o seu valor depende ou não da ocorrência de certos fenômenos.

Os valores que são apontados como potenciais causas da variação da variável dependente (aplicação de uma vacina, utilização de um remédio, características dos autores) são chamados de variáveis independentes. Não se busca explicar a causa dessas variáveis independentes:

- por que a vacina foi aplicada?
- por que um remédio foi usado?
- por que certos autores moveram determinadas ações?

Essas são perguntas que não interessam à pesquisa sobre as causas da variável dependente, embora possam interessar a outras pesquisas: abordagens históricas sobre desenvolvimentos de vacinas, abordagens sociológicas sobre remédios que são preferidos, abordagens políticas sobre o comportamento de certos atores políticos. Em suma:

- **Variável independente** é aquela que representa a potencial **causa** do fenômeno. Ela é independente porque entendemos que ela é uma causa do fenômeno estudado e que, por isso, ela não deve ser uma possível consequência dele. Pode também ser chamada de variável **explanatória** ou **dirigente**. Essa variável é manipulada pela pesquisa experimental, ou observada em condições distintas e pré-definidas na pesquisa observacional, para que se avalie se alterações nessa variável geram efeitos distintos, regulares e observáveis.
- **Variável dependente** é aquela que representa os potenciais **efeitos** dos fenômenos estudados.

1.2 Exercícios

Qual a variável dependente e a independente das seguintes pesquisas?

- Investigar se cortes tendem a tomar decisões diversas em casos de assédio sexual, a depender do gênero dos magistrados que as compõem.
- Investigar se juízes levam mais tempo para decidir questões criminais ou questões civis.
- Investigar se um tribunal é mais propenso a deferir habeas corpus com relação a determinado crime do que outros.

> Respostas: 1. Independente: gênero dos magistrados. Dependente: Decisão em uma ou outra direção. 2. Independente: área do direito. Dependente: média de tempo até a tomada de decisão. 3. Independente: tribunal julga-

dor. Dependente: contagem de deferimentos de habeas corpus.

1.3 A independência das variáveis independentes

É preciso tomar um cuidado extremo no sentido de garantir uma coisa: a variável independente não pode depender (no sentido de ser uma possível consequência) da variável dependente. A falta de cuidado nesse ponto pode levar a pesquisas com interpretações distorcidas de certas correlações.

Podemos, por exemplo, partir da observação de que parece haver uma correlação entre a determinação de lockdown e a existência de grandes índices de contaminação de covid. De fato, parece que esses fenômenos estão correlacionados, pois uma das estratégias de enfrentamento de uma epidemia que se alastra é decretar um lockdown.

Localizar essa correlação não é um problema. O problema é que certos pesquisadores, sem cuidado ou com má-fé, podem se questionar: será que o lockdown causa aumento nos casos de covid? Essa inversão na pergunta gera uma abordagem enviesada, pois a eventual correlação existente entre a variação dessas duas medidas (ocorrência de lockdown e maior incidência de casos) pode conduzir a pesquisas que afirmam uma causalidade inexistente.

A situação pode ser ainda pior: certos pesquisadores podem tentar relacionar a decretação de lockdown com o aumento de mortes por covid. De fato, essa correlação pode ocorrer, mas ela é esperada, na medida em que se sabe que os efeitos de uma política de lockdown só se fazem sentir na taxa de mortalidade semanas depois da decretação. Como o tempo de incubação e desenvolvimento da Covid é de cerca de 2 semanas, as pessoas contaminadas antes do lockdown continuarão a ser hospitalizadas e a falecer depois do início do isolamento social forçado.

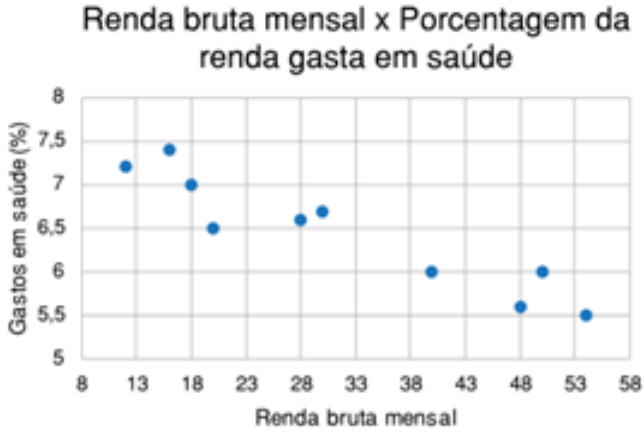
Embora haja correlação e haja também uma sucessão temporal entre os fenômenos, seria no mínimo apressada a conclusão de que a correlação sucessiva é índice de causalidade.

Esses exemplos indicam que é preciso tomar bastante cuidado ao criar as hipóteses de trabalho, pois uma pesquisa pode ter cálculos estatísticos corretos, mas gerar resultados equivocados porque as análises estatísticas pressupõem que as variáveis dependentes não são causadas pelas variáveis independentes.

Esse tipo de risco é um dos motivos pelos quais o conhecimento material é um dos pilares da data science: habilidades estatísticas e computacionais descoladas de um domínio material das temáticas estudadas podem conduzir à formulação de hipóteses enviesadas, que podem conduzir a resultados distorcidos.

2. Gráfico de dispersão

Uma das formas mais claras de visualizar a relação entre duas variáveis (especialmente de duas variáveis numéricas) é através do gráfico de dispersão (scatterplot). Nesse caso, a variável dirigente é geralmente descrita pelo eixo x, e a dirigida pelo eixo y: observe que faz sentido supor que a Renda Bruta Mensal impacte nos percentual gasto em saúde por uma de uma família, mas não faz sentido supor o inverso: que os gastos percentuais em saúde são uma potencial causa da renda bruta mensal.

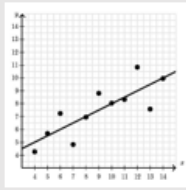


Este gráfico permite visualizar facilmente que existe uma tendência a, quanto maior for a renda familiar, menor ser o percentual de gastos com saúde. Mesmo que, em alguns casos, uma família de maior renda possa dedicar um percentual maior da renda familiar para saúde que outra família de menor renda, isso ocorre em proporções pequenas frente ao todo. O gráfico de dispersão permite visualizar essas proporções.

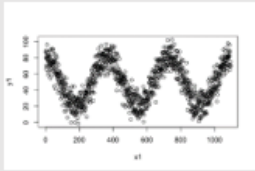
Veja que, quando o eixo x aumenta de valor, o eixo y tende a reduzir de valor. Essa relação entre variáveis é chamada de **correlação**.

A correlação entre duas variáveis é descrita por três fatores: sua linearidade, força e direção.

Linearidade: a correlação entre variáveis pode ser representável por uma reta (linear) ou seguir algum outro padrão (parábolas, ondas, etc.).

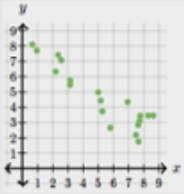


Correlação linear

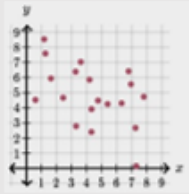


Correlação não linear

Força: a correlação entre duas variáveis pode ser forte, quando uma alteração no valor de x corresponde quase certamente a uma alteração previsível no valor de y; ou fraca, quando uma alteração em x não corresponde a uma alteração claramente previsível no valor de y.

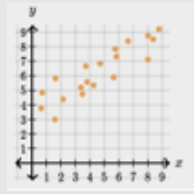


Correlação forte

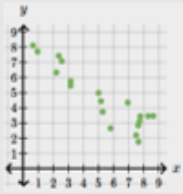


Correlação fraca

Direção: a correlação pode ser **positiva**, quando um aumento do valor de x corresponde a um aumento do valor de y; ou **negativa**, quando um aumento do valor de x gera uma diminuição no valor de y.



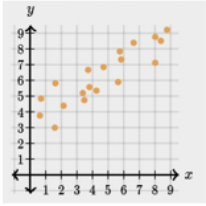
Correlação positiva



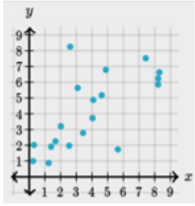
Correlação negativa

Para se avaliar quão forte é a correlação entre duas variáveis (ou quão bem a correlação pode ser descrita por uma linha), estatísticos desenvolveram um coeficiente numérico, o **coeficiente de correlação r**, ou coeficiente de Pearson. Quanto mais próximo de 1, maior a força de uma correlação positiva. Quanto mais próximo de -1, maior a força de uma correlação negativa. Quanto mais próximo de 0, menor é a força da correlação.

Uma ideia intuitiva de _r_é perceptível comparando-se os gráficos de dispersão abaixo. Enquanto ambos os gráficos possuem um índice maior que zero, indicando que existe alguma correlação positiva, o gráfico **A** possui um r maior, o que designa uma correlação mais forte.

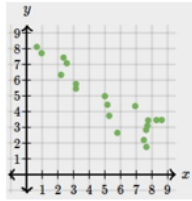


r = 0,89

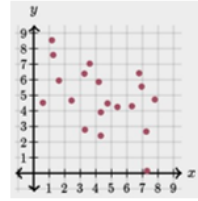


r = 0,67

Os gráficos a seguir descrevem uma correlação bivariada negativa, mas suas forças também são distintas.

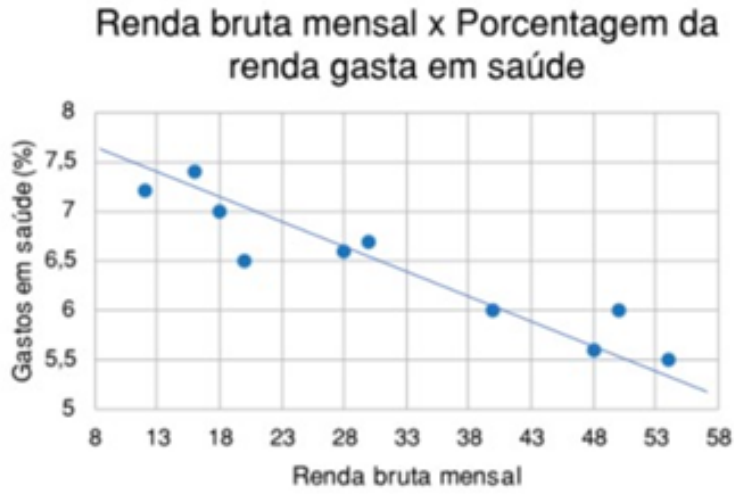


r = -0,92



r = -0,48

Para facilitar a visualização da correlação entre o eixo x e o eixo y, é possível calcular uma linha (ou uma curva) que represente a tendência geral de impacto de uma variável em outra.

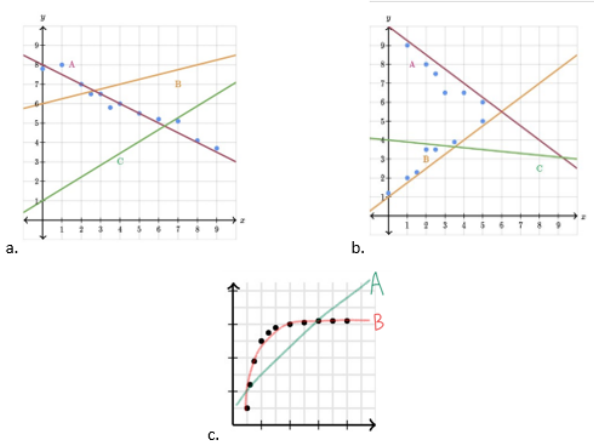


Para que essa linha seja a mais fiel possível frente ao dado observado, deve-se calculá-la de forma que ela seja a linha, dentre todas as linhas possíveis, que esteja à menor distância possível, na média, de cada um dos pontos representados. Essa linha, chamada de **linha de regressão** ou **linha de mínimos quadrados**[4][1].

[1] https://dsd.arcos.org.br/estatistica-instrumental-analise-bivariada/#_ftn4

Você não precisa aprender a calcular os pontos dessa reta, pois esse traçado pode ser feito por programas que você já domina, como o Tableau. O importante é entender o que podemos concluir a partir da observação dessa linha e de suas descrições matemáticas.

Mas primeiramente, vamos praticar a noção intuitiva do desenho da linha de tendência. Qual dessas linhas de tendência melhor se adequa aos dados plotados em cada um dos gráficos a seguir?



Respostas: (a.) R: Linha A; (b.) R: nenhuma linha se adequa aos dados; (c.) R: B, pois a correlação é não-linear.

O coeficiente de correlação nos informa que, para valores próximos de 1 ou -1, a reta será confiável para se prever um valor de y com base em um valor x. Quando r é relativamente próximo de 1 ou -1, é possível então descrever-se a reta matematicamente para se prever o valor de y.

$$y = mx + b$$

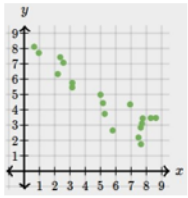
Essa equação, que descreve retas em geral, nos diz que determinado valor de y será igual ao valor correspondente de x multiplicado por

um índice m (a inclinação da reta, ou seja, quanto de y se altera a cada alteração de x), adicionado a um valor fixo b, equivalente ao ponto em que a reta cruza o eixo y. A inclinação será positiva numa correlação positiva e vice-versa.

Prever o valor de uma variável com base em outra variável é o que chamamos de **regressão**. A regressão mais simples é a bivariada linear, mas é possível calcular uma série de regressões mais complexas para correlações não lineares e para múltiplas variáveis.

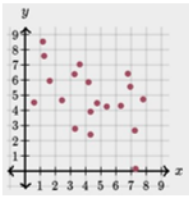
É importante aqui não confundir a inclinação da reta com a força da correlação. A correlação pode ser forte, mas a inclinação da reta ser mínima. Por exemplo, a correlação pode ser extremamente forte (r = 0,98), mas cada aumento (ou diminuição) forte do valor de x gerar um aumento discreto do valor de y.

Ao elevar o valor de r ao quadrado (**r^2**), temos como resultado um outro coeficiente relevante, o **coeficiente de determinação**. Quanto maior o valor de r^2 (que é sempre positivo e entre 0 e 1), melhor a linha consegue explicar a correlação das variáveis.



r = -0,92

r^2 = 0,85

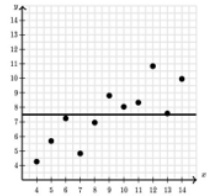


r = -0,48

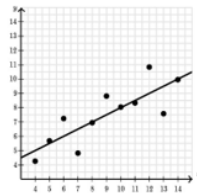
r^2 = 0,23

O cálculo de r^2 é também complexo, mas aqui basta entender um dado fundamental sobre este coeficiente: ele nos diz **quanto de erro de previsão foi eliminado ao se comparar a linha de regressão de y em função de x, em comparação à previsão do valor de y sem nenhuma consideração de x** (ou seja, apenas com base na

média geral de y, ou $\bar{y}$, que é o melhor valor para resumir os pontos em y).



Previsão com base em ȳ



Previsão com base na linha de regressão

Trata-se de uma comparação entre os dados reais observados na amostra com, por um lado, os dados previstos pela linha de regressão, e, de outro, os dados previstos pela linha média do eixo y. Se r^2 é 0,99, isso significa que 99% dos dados são melhor descritos pela linha de regressão do que pela média geral de y (que não depende de x).

Se r^2 é 0,23, então 23% dos pontos são descritos com um erro menor através da linha de regressão, o que significa que 77%[8][2] dos pontos da amostra são descritos com o mesmo grau de eficiência tanto pela linha de regressão quanto pela linha média de y. Quando isso acontece, não faz muita diferença fazer análises usando a linha de regressão ou a média, que é mais fácil de calcular e mais intuitiva. Os "erros" da linha de regressão são chamados de **resíduos**.

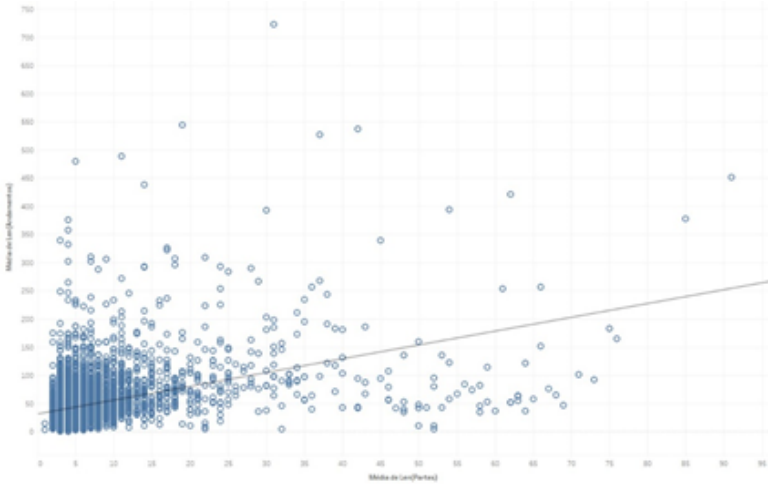
2.1. Aplicação: leitura de dados sobre correlação

O objetivo maior desta introdução em estatística é permitir a leitura de dados que apareçam em artigos ou softwares de processamento

[2] https://dsd.arcos.org.br/estatistica-instrumental-analise-bivariada/#_ftn8

de dados. Vamos então começar entendendo as notações mais comuns que aparecem nesses meios.

No Tableau, a linha de tendência é incluída no gráfico ao se clicar com o botão direito em qualquer espaço vazio do gráfico, e em seguida selecionando-se "Linhas de tendência" e "Mostrar linhas de tendência". Vamos trabalhar com o exemplo abaixo, em que cruzamos o número de partes em cada processo (x) e a média de andamentos, para avaliar se processos com mais parte tendem a ter mais andamentos.



Ao clicar com o botão direito na própria linha de tendência, e em seguida "descrever linha de tendência", surge esse quadro:

Veja que o quadro nos apresenta a equação da linha de tendência (y = mx + b), que nesse caso apresenta um m (inclinação) de 2,44988 e um b (valor onde a linha cruza o eixo y) de 31,1981.

Informações mais úteis para nossos objetivos nessa introdução vão aparecer ao se voltar para o gráfico, clicar com o botão direito na linha, e selecionar "descrever modelo de tendência".

Veja que esse quadro apresenta outro valor que estamos procurando: o coeficiente de determinação (nosso r^2). Aqui, r^2 é 0,163947. O que esse valor nos diz? Ele nos informa que em torno de 16% dos pontos plotados estão, na média, mais próximos da linha de tendência (uma função de x) que de um valor genérico de y que não depende de x (a média geral de y). Dito de outra forma, 84% dos pontos não são melhor explicados pela linha de tendência.

Esse índice de r^2 é extremamente baixo, o que permite ao pesquisador chegar numa conclusão estatisticamente interessante: que não existe uma correlação relevante entre essas variáveis.

Correlação não é causalidade!

Além disso, a correlação pode ser mera coincidência. Por exemplo, existe uma forte correlação (r=0,95) entre o número anual de

formandos em engenharia civil nos EUA e o consumo per capita de queijo muçarela naquele mesmo país, mas seria absurdo sustentar que exista qualquer relação de causalidade entre as duas variáveis. Esse site traz outros exemplos fascinantes de correlações espúrias, ou seja, fruto de coincidências: https://www.tylervigen.com/spurious-correlations[3]

---

[1][4] Curvas de distribuição, ou curvas de densidade, são representações gráficas, em um plano cartesiano, da distribuição de um conjunto de dados em um contínuo (sem intervalos entre os valores).

[2][5]https://galtonboard.com/probabilityexamplesinlife

[4][6] Quadrados pois as distâncias são medidas elevadas ao quadrado, para que pouco importe se elas forem negativas ou positivas em relação ao ponto central, e para que outliers tenham um peso exponencialmente maior na medição do erro de previsão.

[5][7] R: Linha A.

[6][8] R: nenhuma linha se adequa aos dados.

[7][9] R: B, pois a correlação é não-linear.

[8][10] 0,77 = 1 - 0,23

[3]https://www.tylervigen.com/spurious-correlations
[4]https://dsd.arcos.org.br/estatistica-instrumental-analise-bivariada/#_ftnref1
[5]https://dsd.arcos.org.br/estatistica-instrumental-analise-bivariada/#_ftnref2
[6]https://dsd.arcos.org.br/estatistica-instrumental-analise-bivariada/#_ftnref4
[7]https://dsd.arcos.org.br/estatistica-instrumental-analise-bivariada/#_ftnref5
[8]https://dsd.arcos.org.br/estatistica-instrumental-analise-bivariada/#_ftnref6
[9]https://dsd.arcos.org.br/estatistica-instrumental-analise-bivariada/#_ftnref7
[10]https://dsd.arcos.org.br/estatistica-instrumental-analise-bivariada/#_ftnref8