VII BIENAL DA SOCIEDADE BRASILEIRA DE MATEMÁTICA
UNIVERSIDADE FEDERAL DE ALAGOAS – MACEIÓ – ALAGOAS
2 A 6 DE NOVEMBRO DE 2014

AMOSTRAGEM, REGRESSÃO LINEAR E O MÉTODO DOS MÍNIMOS QUADRADOS EM ESTATÍSTICA: UMA INTRODUÇÃO PARA PROFESSORES DO ENSINO BÁSICO

Humberto José Bortolossi e David da Costa Pinho

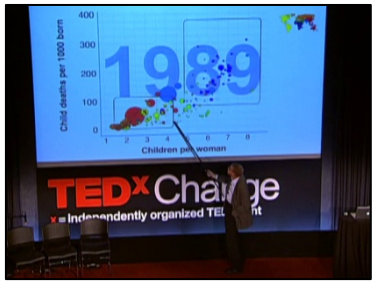
Universidade Federal Fluminense
Instituto GeoGebra Internacional do Rio de Janeiro

CICLO PPDAC

O foco destes cursos [de formação de professores] deve estar nas ideias estatísticas fundamentais, enquanto que, ao mesmo tempo, os professores devem experimentar o ciclo completo da investigação estatística**: problema, planejamento, dados, análise e conclusão** (em Inglês, **PPDAC**: problem, planning, data, analysis, conclusion).

Referência: Batanero, C., Burrill, G., & Reading, C. (2011). *Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education*. A Joint ICMI/IASE Study. ICMI Study volume 14. New York: Springer.

EXEMPLOS DE DAC E PPDAC



Hans Rosling: As Boas Notícias da Década?

http://www.ted.com/talks/lang/pt-br/hans_rosling_the_good_news_of_the_decade.html

HansRosling_2010X-480p.mp4

Hans Rosling: Religiões e Bebês

http://www.ted.com/talks/lang/pt-br/hans_rosling_religions_and_babies.html

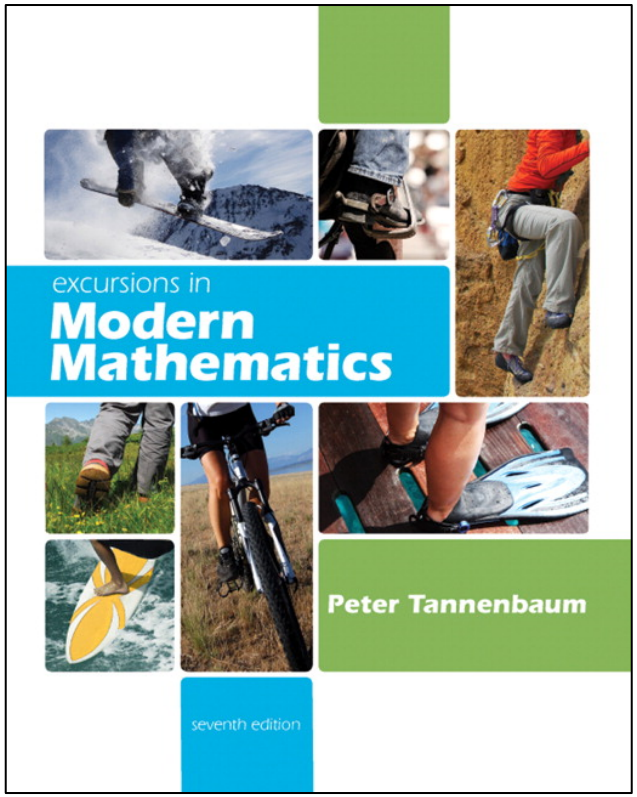
HansRosling_2012S-480p.mp4

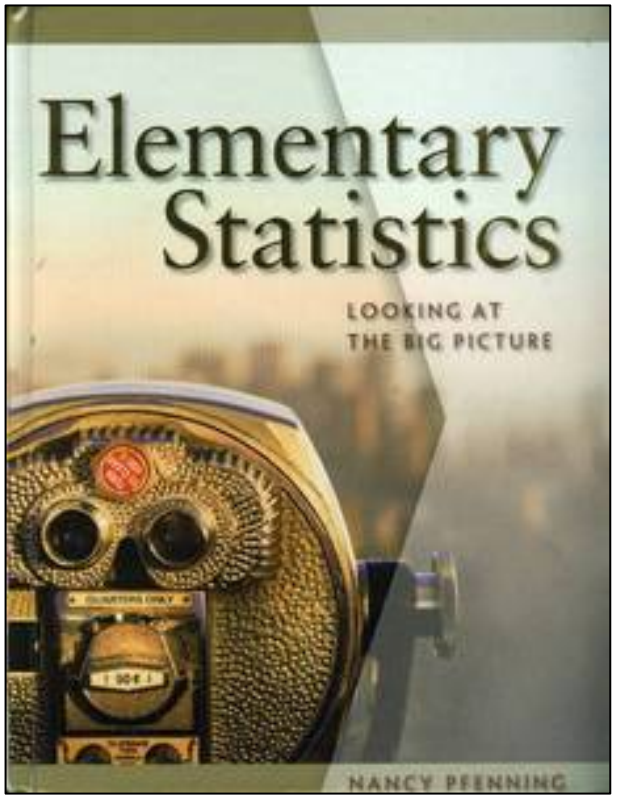
PARTE 1
AMOSTRAGEM: COLETANDO DADOS ESTATÍSTICOS

CONCEITOS-CHAVE

- População, Valor *N* e Censo
- Amostra e Base de Amostragem
- Viés de Seleção, Amostragem por Conveniência, Autosseleção
- Amostragem por Cotas, Aleatória e Estratificada
- Estatística e Parâmetro
- Erro de Amostragem: erros aleatórios e viés de amostragem
- Estudos Clínicos: variáveis de confusão, grupo de tratamento, grupo de controle. Efeito Placebo, Estudo Cego e Estudo Duplo-Cego.

DUAS REFERÊNCIAS PRINCIPAIS





POPULAÇÃO

Toda afirmação estatística se refere, direta ou indiretamente, a algum grupo de indivíduos ou objetos.

Na terminologia estatística, esta coleção de indivíduos ou objetos é denominada **população**.

O primeiro passo para entender uma afirmação estatística é identificar qual é a população a qual ela se refere.

No mundo real nem sempre é fácil identificar a população: detalhes da estória são omitidos ou, alternativamente, duas (ou mais populações) podem estar envolvidas.

EXEMPLO 13.1: O RETORNO DA ÁGUIA AMERICANA

Bald Eagles Come Back from the Brink

BY JOHN L. ELIOT

They ruled the skies on seven-foot (two-meter) wingspans when 17th-century Europeans arrived in North America. Throughout the continent, half a million bald eagles may have soared. But settlers blamed them for killing livestock, so shooting began—and the proud birds' numbers began to plunge. . . .

The Bald Eagle Protection Act of 1940 prohibited shooting or otherwise harming the birds in the [lower 48 states] but didn't cover the pesticides that within a decade began to destroy eagles' eggs. By the 1960s only about 400 breeding pairs of bald eagles remained in the lower 48 [states] . . . . The banning of DDT in 1972 and other measures launched an amazing comeback by the eagles, whose status changed from endangered to threatened in 1995. Today, with more than 6,000 breeding pairs, bald eagles may soon be taken off the endangered species list entirely, their survival as an icon secured—for now.

National Geographic Magazine, July 2002

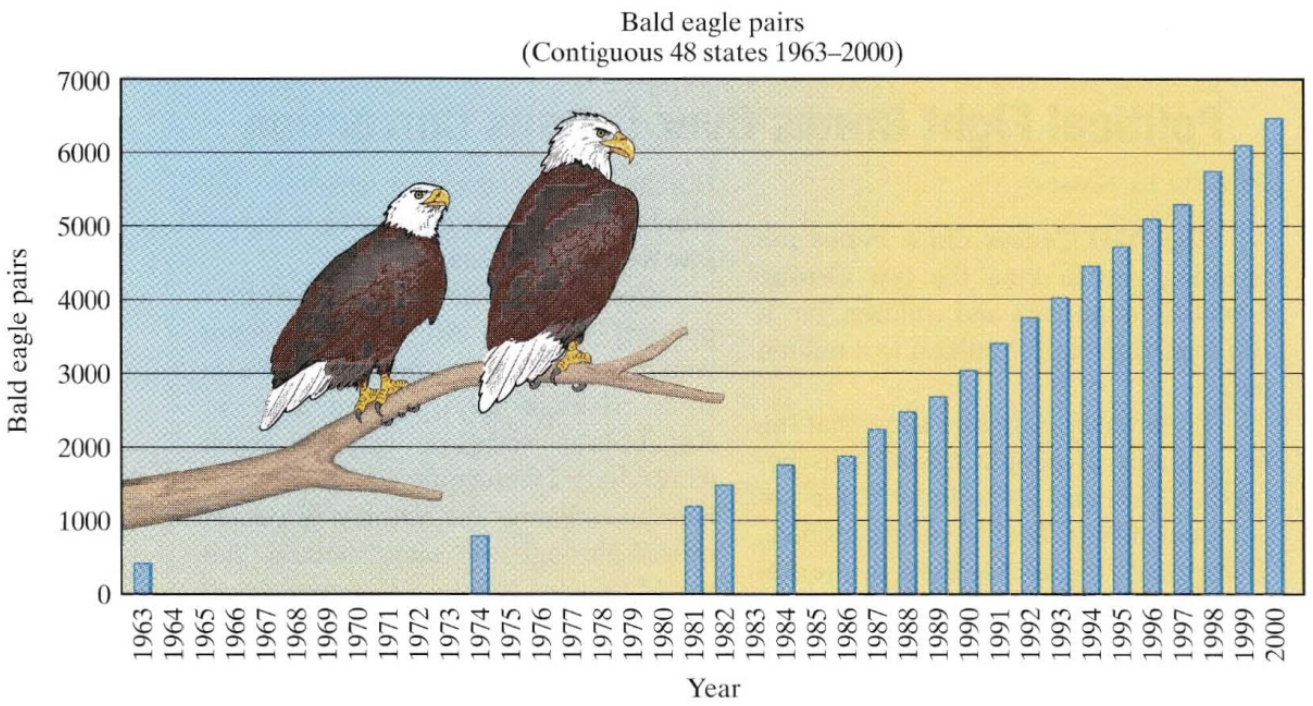
Duas populações: a população total de águias (incluindo filhotes, adolescentes, etc.) nos 48 estados contíguos dos EUA e a população de casais reprodutores de águias.

A primeira é a população de interesse, a segunda é a população de conveniência (mais fácil de se identificar, rastrear e contar). [*From the brink*: da beira do abismo.]

O VALOR-*N*

Valor-*N* (em Inglês, *N-value*) da população é o número de elementos da população.

Importante: ao longo do tempo, uma população e seu valor-*N* podem mudar!



Fonte: Serviço de Pesca e Vida Selvagem dos Estados Unidos.

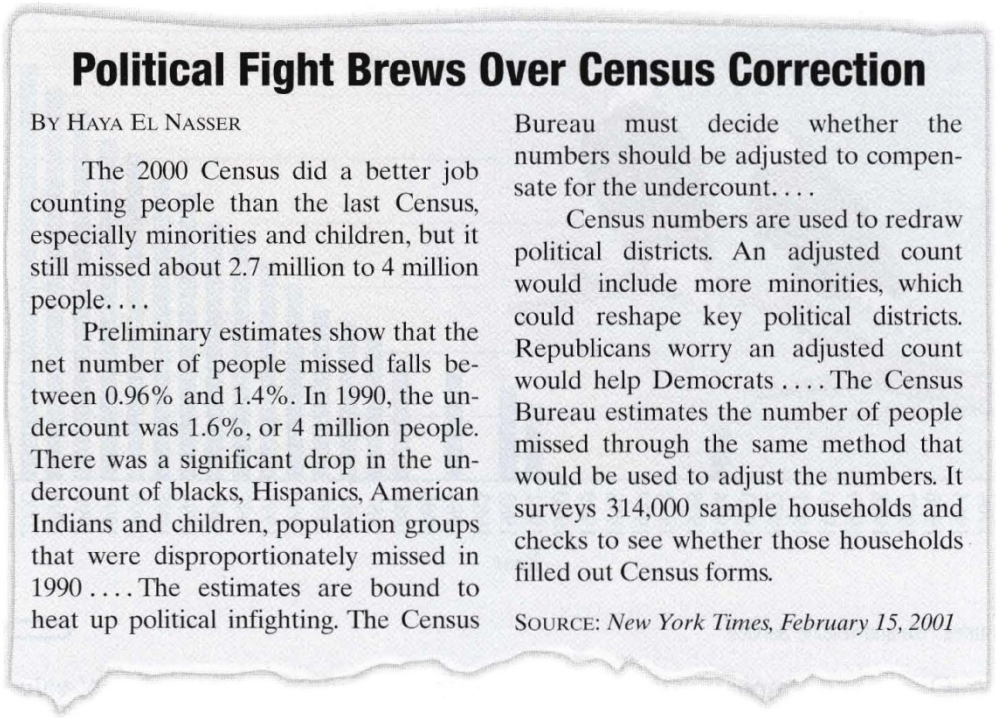
Nenhuma contagem foi realizada em 1964-1973, 1975-1980, 1983 e 1985.

CENSO

Censo (em Inglês, *census*): processo de coletar dados passando por cada membro da população

Exige um alto grau de “cooperação” da população: difícil para populações maiores e mais dinâmicas (vida animal selvagem, humanos, etc.).

EXEMPLO 13.4: O CENSO AMERICANO DE 2000



Political Fight Brews Over Census Correction

BY HAYA EL NASSER

The 2000 Census did a better job counting people than the last Census, especially minorities and children, but it still missed about 2.7 million to 4 million people. . . .

Preliminary estimates show that the net number of people missed falls between 0.96% and 1.4%. In 1990, the undercount was 1.6%, or 4 million people. There was a significant drop in the undercount of blacks, Hispanics, American Indians and children, population groups that were disproportionately missed in 1990 . . . . The estimates are bound to heat up political infighting. The Census Bureau must decide whether the numbers should be adjusted to compensate for the undercount. . . .

Census numbers are used to redraw political districts. An adjusted count would include more minorities, which could reshape key political districts. Republicans worry an adjusted count would help Democrats . . . . The Census Bureau estimates the number of people missed through the same method that would be used to adjust the numbers. It surveys 314,000 sample households and checks to see whether those households filled out Census forms.

SOURCE: *New York Times, February 15, 2001*

O Censo Americano de 2000 empregou cerca de 850 000 pessoas e custou certa de 6,5 bilhões de dólares.

Ainda assim, estima-se que ele deixou de contar entre 3 e 4 milhões de pessoas.

O artigo do New York Times aponta para as implicações políticas desse fato.

ESTUDO DE CASO 1: O CENSO AMERICANO

O Artigo 1 da Seção 2 da Constituição dos Estados Unidos (1787) manda que um censo nacional seja conduzido a cada 10 anos.

O objetivo original do censo era “contar cabeças” com dois propósitos: impostos e representação política.

No texto original, para fim de impostos, índios não deveriam ser taxados e um escravo contaria como 3/5 de uma pessoa livre.

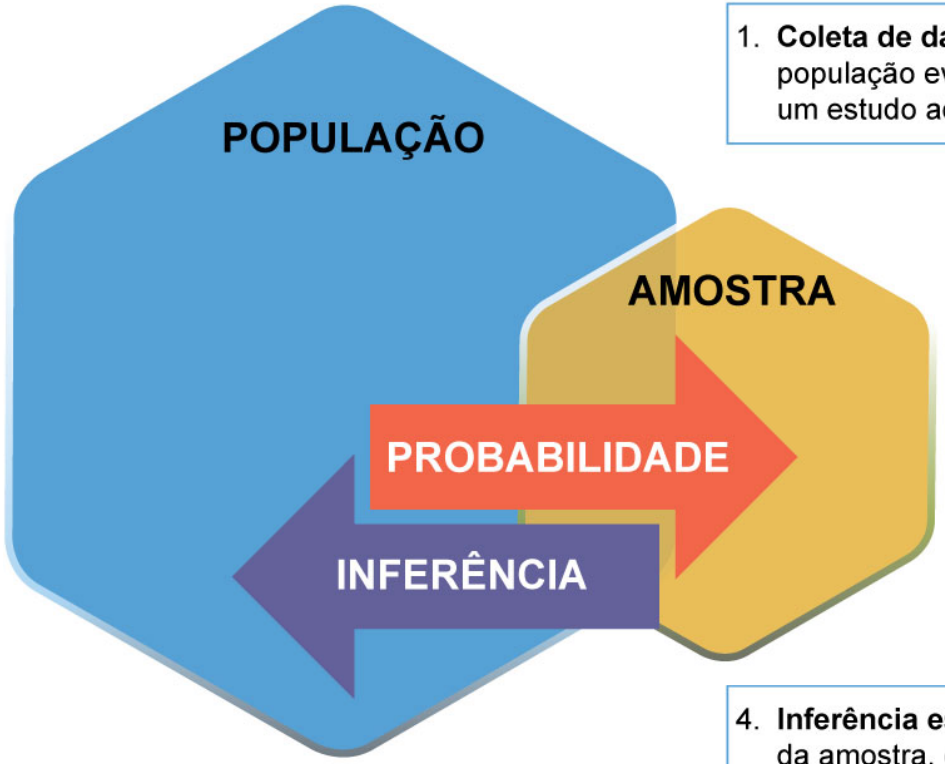
ESTUDO DE CASO 1: O CENSO AMERICANO

O moderno censo dos EUA é atormentado pelo que é conhecido como *subestimação diferencial* (em Inglês, *differential undercount*).

Usando técnicas estatísticas modernas, é possível fazer ajustes nos números brutos do censo que corrigem a imprecisão provocada pela subestimação diferencial.

Caso Departamento de Comércio et al. versus Câmera dos Deputados et al. (1999): a Suprema Corte decidiu no que apenas os números brutos e não os estatisticamente ajustados poderiam ser usados para fins de distribuição de assentos no Congresso entre os estados.

FASES DO PROCESSO ESTATÍSTICO



1. **Coleta de dados**: obter uma amostra representativa da população evitando-se vieses de amostragem através de um estudo adequado.

2. **Desenho e resumo dos dados**: usar diagramas e medidas numéricas dos dados amostrados.

3. **Probabilidade**: conhecendo-se propriedades da população, como amostras aleatórias devem se comportar?

4. **Inferência estatística**: conhecendo-se apenas propriedades da amostra, o que podemos inferir da população como um todo?

SEÇÃO 13.2: AMOSTRAGEM

AMOSTRAGEM

A alternativa prática para um censo é coletar dados somente de alguns membros da população e usar esses dados para obter conclusões e fazer inferências sobre a população inteira.

Esse procedimento é denominado ***survey*** (ou de ***poll*** quando a coleta de dados é feita através de questões). No Brasil, *surveys* e *polls* são denominados genericamente de **pesquisas**.

O subgrupo escolhido que irá fornecer os dados é denominado **amostra** (*sample* em Inglês) e o ato de se selecionar uma amostra é denominado **amostragem** (*sampling* em Inglês).

AMOSTRAGEM

Filosofia básica da amostragem: para obter dados confiáveis, devemos (a) encontrar uma amostra representativa e (b) determinar o tamanho da amostra.

Para populações altamente homogêneas [exemplos?], amostras muito pequenas podem ser usadas.

O primeiro passo importante em uma pesquisa é distinguir a população para a qual a pesquisa se aplica (**população-alvo**) (*target population* em Inglês) e o subconjunto efetivo da população da qual a amostra será tomada, denominado de **base de amostragem** (*sampling frame* em Inglês).

ESTUDO DE CASO 2: A PESQUISA DE OPINÃO PÚBLICA DE 1936 DA LITERARY DIGEST

Landon (republicano) x Roosevelt (democrata)
(presidência dos EUA em 1936)

Resultado da pesquisa da revista:

Landon: 57% Roosevelt: 43%

Base de amostragem usada pela *Literary Digest*: 10 milhões nomes que incluía (1) toda pessoa com nome em alguma **lista telefônica** nos Estados Unidos, (2) toda pessoa que **assinava alguma revista** na época e (3) toda **pessoa inscrita em algum clube ou associação profissional**. [Foram enviadas por correio cédulas de votação fictícias.]

ESTUDO DE CASO 2: A PESQUISA DE OPINÃO PÚBLICA DE 1936 DA LITERARY DIGEST

Landon (republicano) x Roosevelt (democrata)
(presidência dos EUA em 1936)

Resultado da eleição:

Landon: 38% Roosevelt: 62%

George Gallup foi capaz de prever com precisão uma vitória para Roosevelt com uma amostra de “apenas” 50 000 pessoas.

O que deu errado com a enquete *Literary Digest* e por que Gallup conseguiu fazer muito melhor?

ESTUDO DE CASO 2: A PESQUISA DE OPINÃO PÚBLICA DE 1936 DA LITERARY DIGEST

Landon (republicano) x Roosevelt (democrata)
(presidência dos EUA em 1936)

Resultado da eleição:

Landon: 38% Roosevelt: 62%

A base de amostragem não era representativa [lembrar da crise americana da década de 1930].

Após a eleição, a revista perdeu credibilidade e viu suas vendas diminuírem drasticamente, tendo que fechar (uma vítima de um grande erro estatístico).

ESTUDO DE CASO 2: A PESQUISA DE OPINÃO PÚBLICA DE 1936 DA LITERARY DIGEST

Landon (republicano) x Roosevelt (democrata)
(presidência dos EUA em 1936)

Resultado da eleição:

Landon: 38% Roosevelt: 62%

Quando a escolha da amostra tem uma tendência embutida (intencional ou não) para excluir um determinado grupo ou característica na população, dizemos que a pesquisa sofre de um **viés de seleção** (*selection bias* em Inglês).

ESTUDO DE CASO 2: A PESQUISA DE OPINÃO PÚBLICA DE 1936 DA LITERARY DIGEST

Landon (republicano) x Roosevelt (democrata)
(presidência dos EUA em 1936)

Resultado da eleição:

Landon: 38% Roosevelt: 62%

O segundo problema sério com a enquete da *Literary Digest* foi a questão do **viés de não-resposta** (*nonresponse bias*, em Inglês): dos 10 milhões, apenas 2.4 milhões devolveram a cédula preenchida para a revista.

ESTUDO DE CASO 2: A PESQUISA DE OPINÃO PÚBLICA DE 1936 DA LITERARY DIGEST

Landon (republicano) x Roosevelt (democrata)
(presidência dos EUA em 1936)

Resultado da eleição:

Landon: 38% Roosevelt: 62%

Aqueles indivíduos que não querem participar da pesquisa são chamados de **não-respondentes** (*nonrespondents*, em Inglês) e aqueles que participam são chamados de **respondentes** (*respondents*, em Inglês). A percentagem de respondentes na amostra total é chamada de **taxa de resposta** (*response rate*, em Inglês).

ESTUDO DE CASO 2: A PESQUISA DE OPINÃO PÚBLICA DE 1936 DA LITERARY DIGEST

Landon (republicano) x Roosevelt (democrata)
(presidência dos EUA em 1936)

Resultado da eleição:

Landon: 38% Roosevelt: 62%

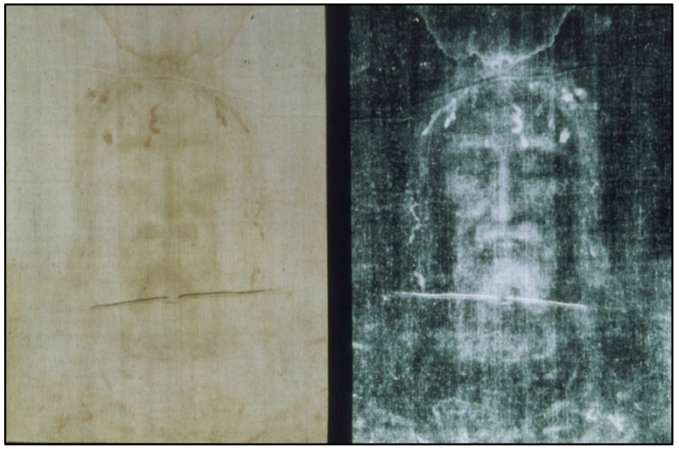
A estória da *Literary Digest* tem duas morais: (1) é melhor usar uma amostra pequena bem escolhida do que uma amostra grande mal escolhida, e (2) fique atento para vieses de seleção e vieses de não-resposta.

AMOSTRAGEM POR CONVENIÊNCIA

Uma técnica comumente usada em amostragem é conhecida como **amostragem por conveniência** (*convenience sampling*, em Inglês). Na amostragem por conveniência, a seleção de quais indivíduos estarão na amostra é feita **seguindo o que é mais fácil ou barato para o coletor de dados**, **sem a preocupação de se obter uma amostra representativa**.

Um exemplo clássico: entrevistadores que pedem para os transeuntes de um local fixo (supermercado, shopping) para participarem de uma pesquisa de opinião pública.

DATAÇÃO POR CARBONO 14 DO SANTO SUDÁRIO

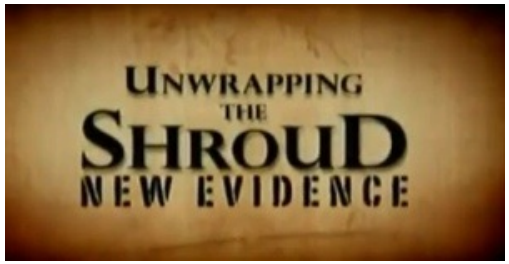


Turin Shroud shown to be a fake

DATAÇÃO POR CARBONO 14 DO SANTO SUDÁRIO

Amostra

Discovery CHANNEL



AMOSTRAGEM POR CONVENIÊNCIA

Um tipo diferente de amostragem por conveniência ocorre quando a amostra é baseada na **autosseleção** (*self-selection*, em Inglês), isto é, a amostra é constituída por indivíduos que se oferecem para estar nela.

Exemplo: enquetes feitas em programas de televisão (que não são confiáveis).

Amostragem por conveniência não é sempre ruim (às vezes, não há outra escolha ou as alternativas são muito caras).

Mas lembre-se: você recebe o que você paga.

O importante é sempre conhecer os detalhes de como os dados foram coletados!

UM EXEMPLO NO BRASIL

Pesquisa feita em 13 de junho de 2013

< http://www.youtube.com/watch?v=6dk0sdyYcdY >

AMOSTRAGEM POR COTAS

A **amostragem por cotas** (*quota sampling*, em Inglês) é um esforço sistemático para forçar que a amostra seja representativa de uma determinada população através do uso de cotas: a amostra deve ter tantas mulheres, tantos homens, tantos negros, tantos brancos, tantas pessoas que vivem em áreas urbanas, tantas pessoas que vivem em áreas rurais, e assim por diante. As proporções de cada categoria na amostra devem ser as mesmas que na população original.

Nosso próximo estudo de caso ilustra algumas das dificuldades com os pressupostos por de trás da amostragem por cotas.

ESTUDO DE CASO 3: A ELEIÇÃO PRESIDENCIAL DOS ESTADOS UNIDOS EM 1948

George Gallup introduziu amostragem por cotas já em 1935 e a usou com sucesso para prever o vencedor das eleições presidenciais nos Estados Unidos em 1936, 1940 e 1944.

A amostra por cotas, portanto, adquiriu a reputação de ser um método de amostragem “cientificamente confiável”.

Em 1948, todas as três principais pesquisas nacionais (Gallup, Roper e Crossley) usaram amostragem por cotas para fazer suas previsões.

ESTUDO DE CASO 3: A ELEIÇÃO PRESIDENCIAL DOS ESTADOS UNIDOS EM 1948

Para a eleição de 1948 entre Thomas Dewey e Harry Truman, Gallup conduziu uma pesquisa com uma amostra de cerca de 3 250 pessoas.

Cada indivíduo na amostra foi entrevistado pessoalmente por um entrevistador profissional para minimizar o viés de não-resposta, e a cada entrevistador foi dado um conjunto muito detalhado de cotas para atender: sete homens brancos com menos de 40 que vivem em uma área rural, 5 homens negros com mais de 40 que vivem em uma área rural, 6 mulheres brancas com menos de 40 que vivem em uma área urbana, e assim por diante.

ESTUDO DE CASO 3: A ELEIÇÃO PRESIDENCIAL DOS ESTADOS UNIDOS EM 1948

Previsão de Gallup:

Dewey: 49,5% Truman 44,5%

Pesquisas Roper e Crossley: também previram uma vitória fácil para Dewey.

Roper, depois de uma pesquisa em setembro mostrando uma diferença de 13 pontos, anunciou que iria descontinuar pesquisas futuras, uma vez que o resultado já era tão óbvio.

ESTUDO DE CASO 3: A ELEIÇÃO PRESIDENCIAL DOS ESTADOS UNIDOS EM 1948

Previsão de Gallup:

Dewey: 49,5% Truman 44,5%

Chicago Daily Tribune

THE WORLD'S GREATEST NEWSPAPER

AN AMERICAN PAPER FOR AMERICANS

HOME

VOL. CVII - NO. 264 WEDNESDAY, NOVEMBER 3, 1948 FOUR CENTS—PAY NO MORE

DEWEY DEFEATS TRUMAN

G.O.P. Sweep Indicated in State; Boyle Leads in City

REPUBLICAN TICKET AHEAD OF 1944 VOTE

Town Balloting Gives Trend

Tops Coghlan in Hot Race for Attorney

Probable Winners

RECORD CITY VOTE SEEN IN LATE TALLIES

Suburban Ballot Near 375,000

BULLETINS ON ELECTIONS

COOK COUNTY

NATION

Early Count Gives G.O.P. Senate Edge

PUTS G.O.P. BACK IN THE WHITE HOUSE

Sizable Electoral Margin Seen

O *Chicago Daily Tribune* estava tão convencido da vitória de Dewey, que em sua primeira edição do dia 04/11/1948 pôs a seguinte manchete: "Dewey defeats Truman".

ESTUDO DE CASO 3: A ELEIÇÃO PRESIDENCIAL DOS ESTADOS UNIDOS EM 1948

Resultado:

Dewey: 44,5% Truman 49,9%

A imagem de Truman erguendo uma cópia do *Tribune* e sua fala de então "*That ain't the way I heard it!*“ tornaram-se parte do folclore norte-americano.

ESTUDO DE CASO 3: A ELEIÇÃO PRESIDENCIAL DOS ESTADOS UNIDOS EM 1948

Para pesquisadores e estatísticos, as previsões erradas da eleição de 1948 deram duas lições: (1) faça pesquisas até o dia da eleição e (2) amostragem por cotas é intrinsecamente falha (até onde vamos parar?).

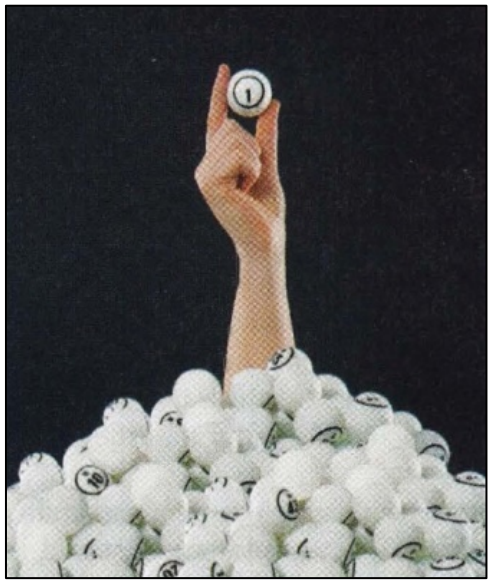
Uma falha ainda mais grave na amostragem por cotas é que, além de satisfazer as cotas, **os entrevistadores estão livres para escolher quem eles querem entrevistar**. Isso abre a porta para um viés de seleção. Olhando para trás ao longo da história da amostragem por cotas nos Estados Unidos, podemos ver uma clara tendência a superestimar o voto republicano.

SEÇÃO 13.3: AMOSTRAGEM ALEATÓRIA

AMOSTRAGEM ALEATÓRIA

Amostragem aleatória (*random sampling* em Inglês): a amostra é obtida por métodos aleatórios.

Princípio: qualquer conjunto de elementos de tamanho *n* tem uma oportunidade igual de ser escolhido como qualquer outro conjunto de mesmo tamanho (*n*).



AMOSTRAGEM ESTRATIFICADA

A amostragem aleatória simples é uma boa ideia para populações pequenas e compactas, mas não é adequada quando se trata de pesquisas nacionais e sondagens de opinião pública.

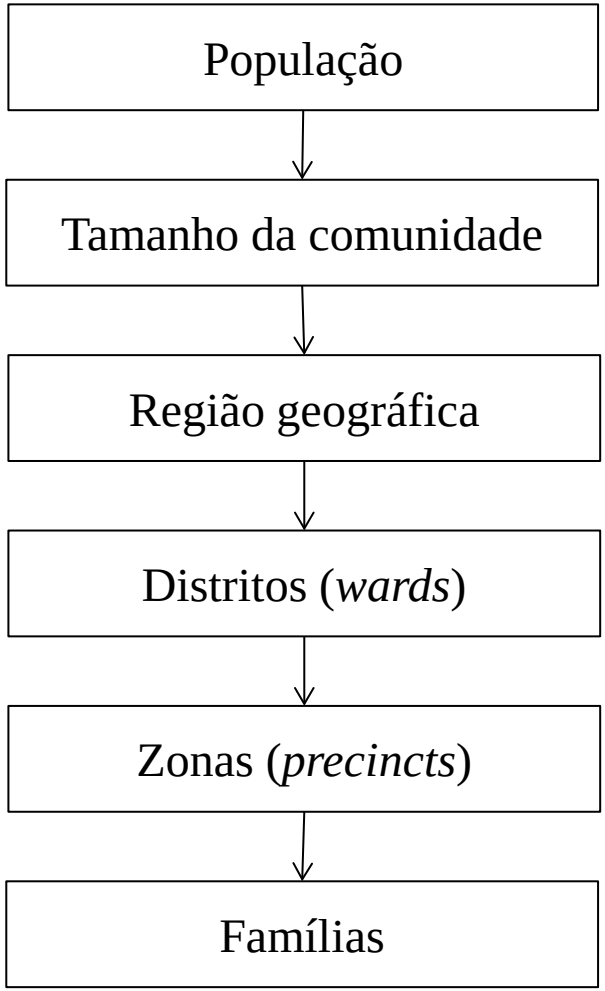
A implementação de amostragem aleatória simples em pesquisas de opinião pública nacional levanta problemas de conveniência e custo.

A alternativa à amostragem aleatória simples usada hoje em dia para sondagens nacionais e pesquisas de opinião pública é um método de amostragem conhecida como **amostragem estratificada** (*stratified sampling*, em Inglês).

AMOSTRAGEM ESTRATIFICADA

A ideia básica da amostragem estratificada é quebrar a base de amostragem em categorias, denominadas **estratos** (*strata*, em Inglês) e, em seguida, **(ao contrário de amostragem por cotas)** escolher **aleatoriamente** uma amostra desses estratos. Os estratos escolhidos são, então, divididos em categorias, denominadas **substratos**, e uma amostra aleatória é tomada destes substratos. Os substratos selecionados são ainda subdivididos, uma amostra aleatória é tomada a partir deles e assim por diante. O processo continua por um número pré-determinado de passos (geralmente quatro ou cinco).

ESTUDO DE CASO 4: PESQUISAS NACIONAIS DE OPINIÃO PÚBLICA



Cidades grandes, cidades médias, cidades pequenas, vilas e áreas rurais

Nova Inglaterra, Atlântico Médio, Estados Montanhosos, etc.

Aos entrevistadores são dadas instruções específicas quanto a quais famílias eles devem realizar entrevistas e a ordem que eles devem seguir

PESQUISAS ELEITORAIS NO BRASIL
DATAFOLHA DE 03/09/2014

Plano amostral e ponderação quanto a sexo, idade, grau de instrução e nível econômico do entrevistado, margem de erro e nível de confiança: Universo: Eleitorado brasileiro, com 16 anos ou mais. Técnica de amostragem: A amostra é estratificada por região geográfica, Unidade da Federação, porte dos municípios e natureza (capital, outros municípios da região metropolitana ou interior) dos municípios. Em cada estrato, num primeiro estágio, são sorteados os municípios que farão parte do levantamento. Num segundo estágio, são sorteados os bairros e pontos de abordagem onde serão aplicadas as entrevistas. Por fim, os entrevistados são selecionados aleatoriamente para responder ao questionário, de acordo com cotas de sexo e faixa etária.

PESQUISAS ELEITORAIS NO BRASIL
DATAFOLHA DE 03/09/2014

Esta pesquisa custou R$ 266.200,00 e entrevistou 10.336 pessoas.

SEÇÃO 13.4: TERMINOLOGIA E CONCEITOS CHAVES EM AMOSTRAGEM

TERMINOLOGIA E CONCEITOS CHAVES EM AMOSTRAGEM

Estatísticos usam o termo **estatística** (*statistics*, em Inglês) para descrever qualquer tipo de informação numérica obtida a partir de uma amostra. Uma estatística é sempre uma estimativa para alguma medida desconhecida, chamada um **parâmetro** (*parameter*, em Inglês) da população.

Erro de amostragem (*sampling error*, em Inglês) (erro amostral): diferença entre um parâmetro e uma estatística utilizada para estimar o parâmetro.

Erros aleatórios (*chance error*, em Inglês) são o resultado do fato básico de que uma amostra, sendo apenas uma amostra, só pode nos dar uma informação aproximada sobre a população.

TERMINOLOGIA E CONCEITOS CHAVES EM AMOSTRAGEM

Variabilidade de amostragem (*sampling variability*, em Inglês) (variabilidade amostral): diferentes amostras são suscetíveis de produzir estatísticas diferentes para uma mesma população, mesmo quando as amostras são escolhidas exatamente da mesma maneira. **[Inevitável!]**

Viés de amostragem (*sample bias*, em Inglês) é o resultado da escolha de uma amostra ruim e é um problema muito mais grave do que um erro aleatório. **[Podem ser eliminados através de métodos adequados de seleção da amostra.]**

Proporção de amostragem: n/N (tamanho da amostra dividido pelo tamanho da população).

AMOSTRAGEM: ANÁLISE DE SITUAÇÃO (EXEMPLO 1)

Situação: Um professor quer obter uma amostra de 6 dos 80 alunos de sua turma para saber o que eles acham sobre o livro didático adotado.

Perguntas: Os métodos de amostragem descritos a seguir são não enviesados? Se não são, que tipo de viés pode estar presente?

1. Pedir para os alunos que queiram opinar sobre o livro para levantarem as mãos.
2. Consultar os seis primeiros alunos que vierem ao horário de atendimento extraclasse do professor.
3. Olhar para a lista de chamada e, sem a ajuda de um gerador de números aleatórios, tirar uma amostra "aleatória" de seis nomes.
4. Atribuir a cada aluno presente na sala de aula um número de 1 em diante e, em seguida, usar um *software* ou uma tabela de números aleatórios para selecionar seis nomes.
5. Tomar uma amostra aleatória da lista de alunos matriculados e enviar por correio ou *e-mail* um questionário a cada um deles.

AMOSTRAGEM: ANÁLISE DE SITUAÇÃO (EXEMPLO 1)

Situação: Um professor quer obter uma amostra de 6 dos 80 alunos de sua turma para saber o que eles acham sobre o livro didático adotado.

Perguntas: Os métodos de amostragem descritos a seguir são não enviesados? Se não são, que tipo de viés pode estar presente?

1. Pedir para os alunos que queiram opinar sobre o livro para levantarem as mãos.

Resposta:

Este método irá gerar uma amostra por **autosseleção**, a qual muito provavelmente favorecerá a inclusão de pessoas com fortes sentimentos positivos ou negativos com relação ao livro sendo, portanto, uma amostra enviesada.

AMOSTRAGEM: ANÁLISE DE SITUAÇÃO (EXEMPLO 1)

Situação: Um professor quer obter uma amostra de 6 dos 80 alunos de sua turma para saber o que eles acham sobre o livro didático adotado.

Perguntas: Os métodos de amostragem descritos a seguir são não enviesados? Se não são, que tipo de viés pode estar presente?

2. Consultar os seis primeiros alunos que vierem ao horário de atendimento extraclasse do professor.

Resposta:

Este método de **amostragem por conveniência** provavelmente não obterá uma amostra representativa, pois, em geral, os estudantes que procuram o horário de atendimento são aqueles com mais dificuldades e estes tendem a achar o livro mais difícil de se entender.

AMOSTRAGEM: ANÁLISE DE SITUAÇÃO (EXEMPLO 1)

Situação: Um professor quer obter uma amostra de 6 dos 80 alunos de sua turma para saber o que eles acham sobre o livro didático adotado.

Perguntas: Os métodos de amostragem descritos a seguir são não enviesados? Se não são, que tipo de viés pode estar presente?

3. Olhar para a lista de chamada e, sem a ajuda de um gerador de números aleatórios, tirar uma amostra "aleatória" de seis nomes.

Resposta:

Se o professor tirar os nomes "da sua cabeça" (uma conveniência para o professor), o resultado será uma amostra obtida sem nenhum planejamento científico (*haphazard sample*) a qual, muito provavelmente, não será uma amostra representativa por causa das tendências conscientes e não conscientes na escolha dos nomes.

AMOSTRAGEM: ANÁLISE DE SITUAÇÃO (EXEMPLO 1)

Situação: Um professor quer obter uma amostra de 6 dos 80 alunos de sua turma para saber o que eles acham sobre o livro didático adotado.

Perguntas: Os métodos de amostragem descritos a seguir são não enviesados? Se não são, que tipo de viés pode estar presente?

4. Atribuir a cada aluno presente na sala de aula um número de 1 em diante e, em seguida, usar um *software* ou uma tabela de números aleatórios para selecionar seis nomes.

Resposta:

O problema em se escolher apenas entre os alunos presentes na sala de aula é que a base de amostragem pode não ser representativa da população: nem todos os alunos matriculados vão às aulas. Alunos ausentes podem ter um sentimento negativo com relação à disciplina em geral (incluindo o livro texto) ou, por outro lado, eles podem achar que não precisam comparecer às aulas, pois o livro é bom o suficiente para lhes ensinar tudo o que eles precisam saber!

AMOSTRAGEM: ANÁLISE DE SITUAÇÃO (EXEMPLO 1)

Situação: Um professor quer obter uma amostra de 6 dos 80 alunos de sua turma para saber o que eles acham sobre o livro didático adotado.

Perguntas: Os métodos de amostragem descritos a seguir são não enviesados? Se não são, que tipo de viés pode estar presente?

5. Tomar uma amostra aleatória da lista de alunos matriculados e enviar por correio ou *e-mail* um questionário a cada um deles.

Resposta:

Esta pode ser a melhor opção do professor. Contudo, existe uma boa chance de que nem todos os 6 alunos responderem e, neste caso, o **viés de não-resposta** pode resultar em uma amostra atípica.

SEÇÃO 13.5: O MÉTODO DE CAPTURA-RECAPTURA

O MÉTODO DE CAPTURA-RECAPTURA

Encontrar o valor-*N*: pode ser extremamente difícil e às vezes impossível. Em muitos casos, uma boa estimativa basta. Essas estimativas podem ser calculadas usando-se métodos de amostragem.

O método mais simples de amostragem para estimar o valor-*N* de uma população é denominado **método de captura e recaptura**. Este método é frequentemente usado por biólogos para estimar o tamanho das populações de animais selvagens, o que explica a escolha incomum do termo “captura-recaptura” no nome do método.

O MÉTODO DE CAPTURA-RECAPTURA

Passo 1. **Capturar (amostrar)**: capturar (escolher) uma amostra de tamanho n_1, marcar (identificar) os animais (objetos, pessoas) da amostra e, então, devolvê-los para a população geral.

Passo 2. **Recapturar (amostrar novamente):** depois de um certo período de tempo, capturar uma nova amostra de tamanho n_2 e contar o número k de indivíduos dessa amostra que foram marcados no Passo 1.

Passo 3. **Estimar:** o valor-N da população pode ser estimado pelo número $n_1\, n_2/k$, isto é, $N \approx n_1\, n_2/k$.

O MÉTODO DE CAPTURA-RECAPTURA

Explicação: se as amostras são representativas, então

$$n_1/N \approx k/n_2$$

(a proporção de indivíduos marcados na população e na amostra são aproximadamente iguais), de modo que

$$N \approx n_1\, n_2/k.$$

EXEMPLO 13.6: PEIXES PEQUENOS EM UM LAGO GRANDE

Um lago artificial grande está abastecido de bagres. Como parte de um projeto de pesquisa, é necessário contar o número de bagres no lago. Um censo está fora de questão (esvaziar o lago), de modo que nossa melhor aposta é o método de captura-recaptura.

Passo 1. Para nossa primeira amostra, vamos capturar um número n_1 pré-determinado de bagres, digamos, $n_1 = 200$. Esses peixes são marcados e, então, liberados de volta ao lago.

EXEMPLO 13.6: PEIXES PEQUENOS EM UM LAGO GRANDE

Passo 2. Após esperar tempo suficiente para que os peixes marcados e liberados se dispersem e se misturem no lago artificial, fazemos uma nova amostragem de n_2 bagres. Não é preciso que n_2 seja igual a n_1, mas é uma boa prática que as duas amostras tenham a mesma ordem de grandeza. Digamos, $n_2 = 150$. Dos 150 bagres da segunda amostra, $k = 21$ estavam marcados (faziam parte da amostra original).

EXEMPLO 13.6: PEIXES PEQUENOS EM UM LAGO GRANDE

Passo 3. Assumindo que as amostras são representativas, temos então que

$$N \approx n_1 n_2/k = 200 \times 150/21 \approx 1428{,}57.$$

Obviamente, o valor 1428,57 não deve ser considerado literalmente, uma vez que N deve ser um número inteiro. Além do mais, é bom lembrar que estamos fazendo uma estimativa. Uma conclusão admissível que existem aproximadamente 1400 bagres no lago.

Assim, pesquei numerosos peixes nas margens do Suez, coloquei-lhes anilhas nas caudas e lancei-os de novo ao mar. Alguns meses mais tarde, nas costas da Síria apanhei alguns peixes com os meus anéis.

Júlio Verne (1828 –1905), Vinte Mil Léguas Submarinas

SEÇÃO 13.6: ESTUDOS CLÍNICOS

ESTUDOS CLÍNICOS

Um tipo diferente de processo de coleta de dados é necessário quando estamos tentando estabelecer conexões entre uma causa e um efeito.

Será que aulas de matemática aumentam as chances de se conseguir um bom emprego? Ser um fumante passivo aumenta significativamente o risco de se desenvolver câncer de pulmão? Será que uma dose diária de aspirina reduz as chances de se ter um ataque cardíaco? Os benefícios da terapia de reposição hormonal para mulheres com mais de 50 superam os riscos?

ESTUDOS CLÍNICOS

Estas perguntas do tipo causa-efeito não podem ser respondidas por meio de uma medição imediata e requerem uma observação durante um período prolongado de tempo.

Além disso, nessas situações, o processo de coleta de dados exige uma participação ativa do experimentador. Além de *observação*, *medição* e *registro*, também existe *tratamento*.

ESTUDOS CLÍNICOS

Quando queremos saber se uma determinada causa X produz um certo efeito Y, uma possibilidade é criar um estudo onde a causa X é gerada e seus efeitos são observados. Se o efeito Y é observado, então é possível que X seja realmente uma causa de Y. Estabelecemos assim uma **associação** (*association*) entre a causa X e o efeito Y.

O problema, no entanto, é a possibilidade de que alguma outra causa Z diferente de X tenha produzido o efeito Y e que X que não tenha nada a ver com isso.

ESTUDOS CLÍNICOS

Só porque estabelecemos uma associação, isso não significa que tenhamos estabelecido uma relação de causa-e-efeito entre as variáveis. Estatísticos gostam de explicar isso com um simples ditado: **associação não é causalidade** (*association* is *not causation).*

ESTUDO DE CASO 5: O CASO ALAR

Alar: até 1989, produto químico muito usado por produtores de maçã nos EUA para regular o ritmo em que as maçãs amadurecem.

Em 1989, o Alar é denunciado nos jornais e na TV como um potente agente causador de câncer e uma das principais causas de câncer em crianças.

Resultado: as pessoas pararam de comprar maçãs, escolas de todo o país removeram suco de maçã de seus menus, e a indústria de maças do Estado de Washington perdeu cerca de 375 milhões de dólares.

ESTUDO DE CASO 5: O CASO ALAR

O caso contra o Alar baseou-se em um único estudo em 1973 com ratos de laboratório. A dosagem utilizada no estudo foi oito vezes maior do que a concentração máxima tolerada: uma dosagem em que até mesmo substâncias inofensivas podem produzir danos teciduais.

Uma criança teria que comer cerca de 200 000 maçãs por dia para ser exposta a uma dosagem equivalente.

Estudos posteriores feitos pelo Instituto Nacional do Câncer e pela Agência de Proteção Ambiental falharam em mostrar qualquer relação causa-e-efeito entre o Alar e o câncer em crianças.

ESTUDOS CLÍNICOS

Um **estudo clínico** (*clinical study*) ou **ensaio clínico** (*clinical trial*) é um tipo de estudo preocupado em determinar se uma única variável ou tratamento (normalmente uma vacina, um medicamento, uma terapia, etc.) pode causar um certo efeito (uma doença, um sintoma, uma cura, etc.) .

O primeiro e o mais importante quesito em qualquer estudo clínico é o de se isolar a causa (tratamento, medicamentos, vacinas, terapia, etc.) que está sob investigação de todas as outras possíveis causas (chamadas de **variáveis de confusão** (*confounding variables*) que poderiam produzir o mesmo efeito. Geralmente, isto é feito *controlando-se* o estudo.

ESTUDOS CLÍNICOS

Exemplo de Variáveis de Confusão

Nível de Barulho × Nível de Concentração

O nível de concentração está associado apenas com o nível de barulho?

Variáveis	Condição 1	Condição 2
Nível de Barulho	Baixo	Alto
QI	Médio	Médio
Temperatura da Sala	28ºC	28ºC
Sexo	60% mulheres	60% mulheres
Dificuldade do teste	Moderado	Moderado
Horário do dia	Manhã	Manhã

ESTUDOS CLÍNICOS

Exemplo de Variáveis de Confusão

Nível de Barulho × Nível de Concentração

O nível de concentração está associado apenas com o nível de barulho?

Variáveis	Condição 1	Condição 2
Nível de Barulho	Baixo	Alto
QI	Médio	Médio
Temperatura da Sala	28ºC	40ºC
Sexo	60% mulheres	60% mulheres
Dificuldade do teste	Moderado	Moderado
Horário do dia	Manhã	Tarde

ESTUDOS CLÍNICOS

Em um **estudo controlado** (*controlled study*), os indivíduos são divididos em dois grupos: o *grupo de tratamento* (*treatment group*) e o *grupo de controle* (*control group*).

O **grupo de tratamento** é constituído por aqueles indivíduos que receberam o tratamento de fato. O **grupo de controle** consiste dos indivíduos que **não receberam qualquer tratamento** e que estão participando do estudo apenas para fins comparativos (o grupo controle também é chamado de grupo de *comparação* (*comparison group*)).

Se uma relação real de causa-e-efeito existe entre o tratamento e o efeito a ser estudado, então o grupo de tratamento deve mostrar os efeitos do tratamento e o grupo de controle não.

ESTUDOS CLÍNICOS

Estudo controlado aleatório (*randomized controlled study*): os indivíduos são designados ou para o grupo de tratamento ou para o grupo de controle de forma aleatória.

Efeito placebo (*placebo effect*): princípio geralmente aceito de que *apenas a ideia de que se está recebendo um tratamento pode produzir resultados positivos*.

Estudo cego (*blind study*): nem os membros do grupo de tratamento e os do grupo de controle sabem em qual grupo eles estão.

Estudo duplo-cego (*double blind-study*): nem os participantes e nem os cientistas que conduzem a experiência sabem quais são os indivíduos que estão no grupo de tratamento e quais estão no grupo de controle.

ESTUDO DE CASO 6: OS ENSAIOS DE SALK DE 1954 PARA A VACINA DA PÓLIO

ESTUDO DE CASO 6: OS ENSAIOS DE SALK DE 1954 PARA A VACINA DA PÓLIO

	Número de crianças	Número de casos relatados de pólio	Número de casos com paralisia de pólio	Número de casos fatais de pólio
Grupo de tratamento	200745	82	33	0
Grupo de controle	201229	162	115	4
Se recusaram a participar	338778	182*	121*	0*
Desistiram	8484	2*	1*	0*
Total	749236	428	270	4

*Estes números não são indicadores confiáveis do número real de casos: eles são apenas casos autorrelatados.

Fonte: adaptado de Thomas Francis, Jr., et al., "An Evaluation of The 1954 Poliomyelitis Vaccine Trials – Summary Report", *American Journal of Public Health*, 45 (1955), 25.

Estes dados dão evidências conclusivas de que a vacina Salk foi um tratamento efetivo para a pólio e, com base nesse estudo, uma campanha de inoculação em massa foi estabelecida. Hoje, todas as crianças são inoculadas contra a pólio e a doença foi essencialmente erradicada dos Estados Unidos. A Estatística desempenhou um papel importante neste marco importante em saúde pública.

PARTE 2:
ASSOCIAÇÃO, CORRELAÇÃO E REGRESSÃO LINEAR

MOTIVAÇÃO

Para tentar entender o mundo que nos cerca, uma questão que surge naturalmente é investigar se determinados fenômenos estão associados ou não. Alguns exemplos:

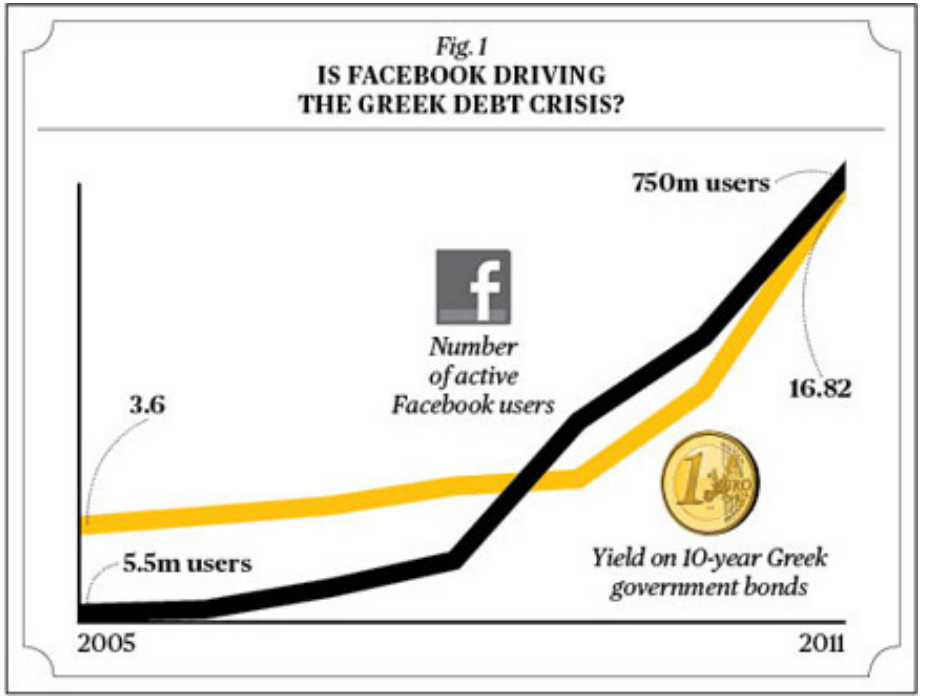
- A probabilidade de se desenvolver um câncer de pulmão está associada com a quantidade de cigarros consumidos?
- O tempo necessário para uma bola metálica maciça atingir o solo está associado com a altura em que ela é largada?
- A medida de uma temperatura em graus Fahrenheit está associada com a medida da mesma temperatura em graus Celsius?

Em Estatística, a **análise de regressão** é um método usado para se estudar tais associações.

MOTIVAÇÃO

Tipicamente, os fenômenos são descritos **variáveis quantitativas** e, então, estuda-se se estas variáveis estão relacionadas (**correlação**) e, se este for o caso, qual é o tipo de relação funcional entre elas (**regressão**).

Correlação não implica causalidade!



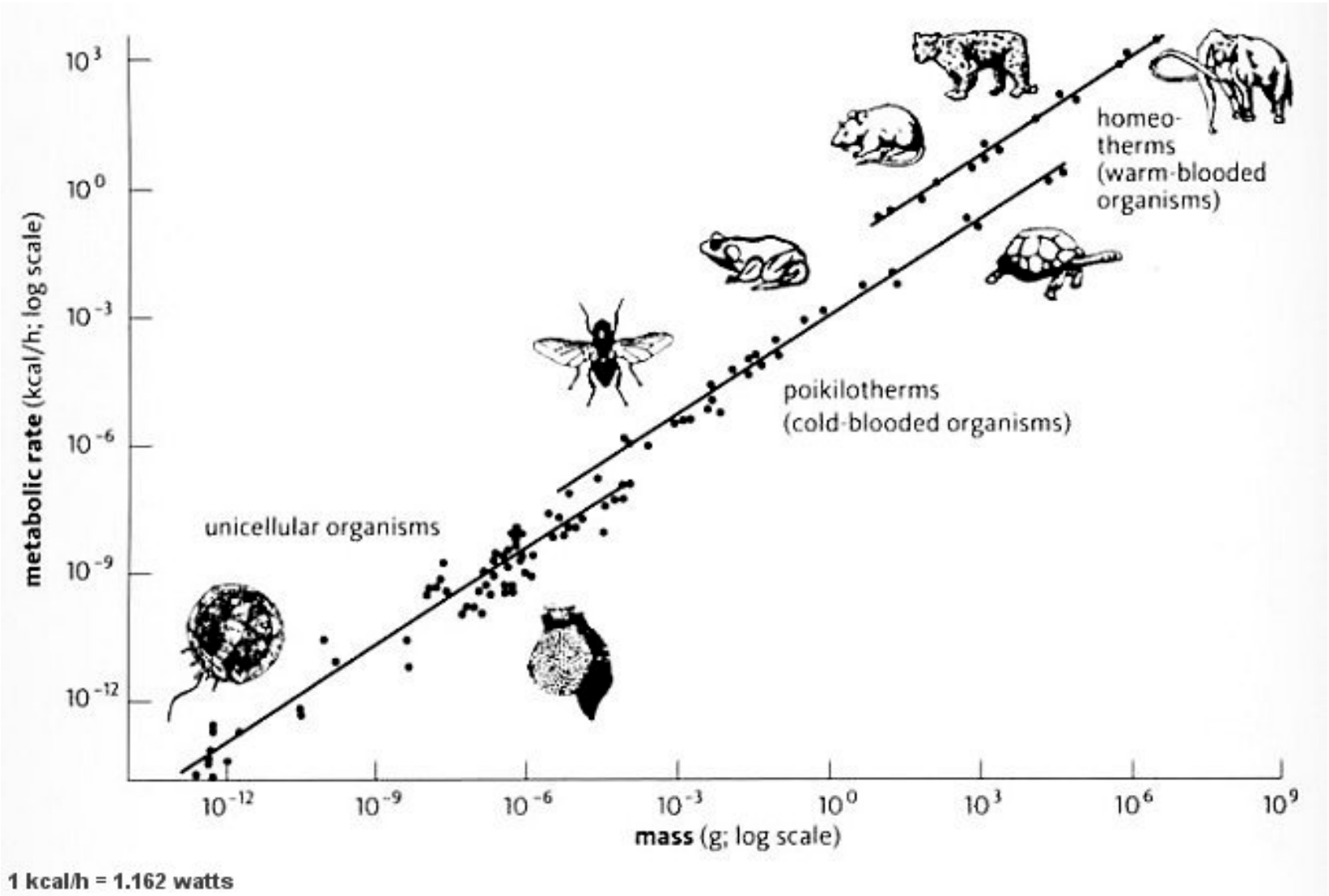
MOTIVAÇÃO

Tipicamente, os fenômenos são descritos **variáveis quantitativas** e, então, estuda-se se estas variáveis estão relacionadas (**correlação**) e, se este for o caso, qual é o tipo de relação funcional entre elas (**regressão**).

Correlação não implica causalidade!

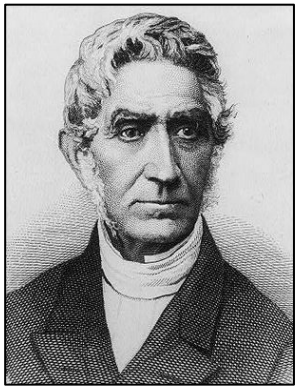
O que as correlações não substituem é o pensamento humano. Você tem que pensar sobre o que aquilo significa. O que um bom cientista faz, se ele chega a uma correlação, é tentar refutá-la o quanto for possível, tentar desconstrui-la, se livrar dela, tentar contestá-la. Se ela aguentar todos esses esforços para destruí-la e ainda estiver de pé, então, cautelosamente, você diz: podemos mesmo ter algo aqui!

Sir Michael Marmot em The Joy of Stats

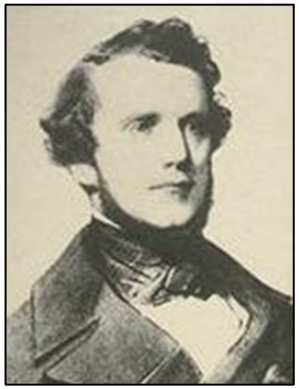


1 kcal/h = 1.162 watts

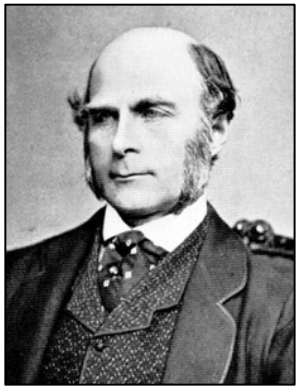
UM BREVE RESUMO HISTÓRICO



Quetelet
(1796-1874)



Bravais
(1811-1863)



Galton
(1822-1911)

Pearson
(1857-1936)

Legendre
(1752-1833)

CORRELAÇÃO E REGRESSÃO NO ENSINO BÁSICO

Para a área de **Matemática**: os PCN não abordam correlação e regressão. O enfoque é em **Estatística Univariada**.

Por quê? Talvez pela Matemática envolvida (Cálculo, Álgebra Linear, ...).

Para outras áreas:

Biologia: [...] elaborar tabelas ou gráficos mostrando a **correlação** entre certos indicadores como mortalidade infantil e escolaridade dos pais, ou níveis de renda e incidência de doenças infecto-contagiosas [...]

Química: **Correlacionar** dados relativos à concentração de certas soluções nos sistemas naturais a possíveis problemas ambientais.

CORRELAÇÃO E REGRESSÃO NO ENSINO BÁSICO

Dedução da fórmula dos coeficientes da reta de regressão linear pelo critério dos mínimos quadrados usando apenas **funções quadráticas**, acessível a alunos do Ensino Médio (uma adaptação de Casella & Berger (2002)).

Por motivos didáticos, apresentaremos aqui a dedução em duas versões:

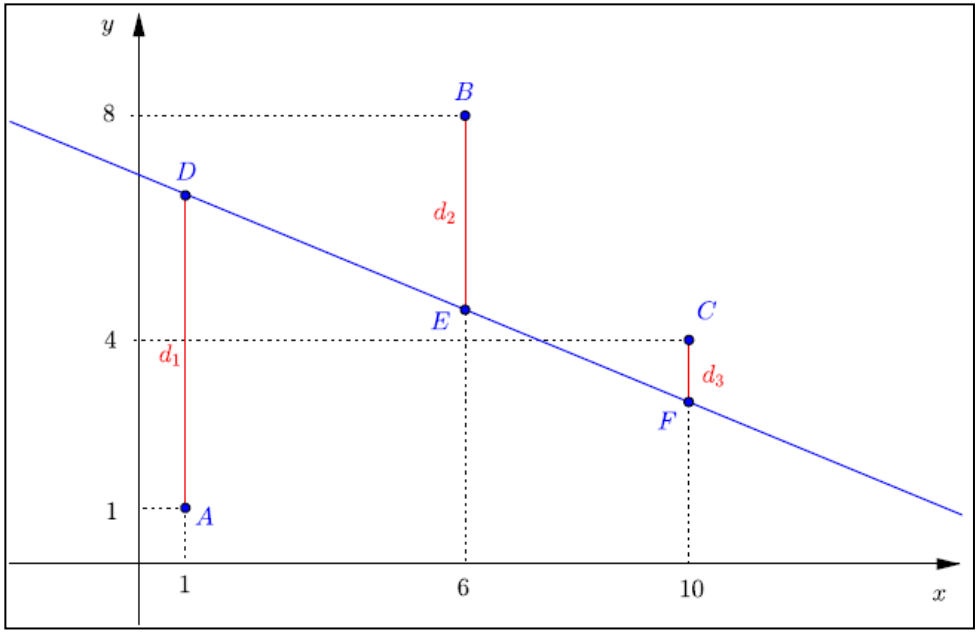
(1) para uma escolha particular de três pontos no plano cartesiano;

(2) para *n* pontos quaisquer no plano cartesiano.

DEDUÇÃO USANDO FUNÇÕES QUADRÁTICAS

Sejam $(x_1, y_1) = (1, 1)$, $(x_2, y_2) = (6, 8)$ e $(x_3, y_3) = (10, 4)$ três pontos no plano cartesiano. Queremos minimizar em a e b (os coeficientes da **reta** $y = a\,x + b$):

$$d_1^2 + d_2^2 + d_3^2 \;=\; (1-(1a+b))^2 + (8-(6a+b))^2 + (4-(10a+b))^2$$



DEDUÇÃO USANDO FUNÇÕES QUADRÁTICAS

Para cada valor de a (isto é, considerando-se a como um parâmetro), considere a função

$$f(b)$$

$$\|$$

$$d_1^2 + d_2^2 + d_3^2$$

$$\|$$

$$(1-(1a+b))^2 + (8-(6a+b))^2 + (4-(10a+b))^2$$

$$\|$$

$$((1-a)-b)^2 + ((8-6a)-b)^2 + ((4-10a)-b)^2$$

$$\|$$

$$(1-a)^2 - 2(1-a)b + b^2 + (8-6a)^2 - 2(8-6a)b + b^2 + (4-10a)^2 - 2(4-10a)b + b^2$$

DEDUÇÃO USANDO FUNÇÕES QUADRÁTICAS

$$f(b)$$

$$\|$$

$$(1-a)^2-2(1-a)b+b^2+(8-6a)^2-2(8-6a)b+b^2+(4-10a)^2-2(4-10a)b+b^2$$

$$\|$$

$$(1-a)^2+(2a-2)b+b^2+(8-6a)^2+(12a-16)b+b^2+(4-10a)^2+(20a-8)b+b^2$$

$$\|$$

$$3b^2+(2a-2+12a-16+20a-8)b+(1-a)^2+(8-6a)^2+(4-10a)^2$$

$$\|$$

$$3b^2+(34a-26)b+(1-a)^2+(8-6a)^2+(4-10a)^2.$$

DEDUÇÃO USANDO FUNÇÕES QUADRÁTICAS

$$f(b) = 3b^2 + (34a - 26)b + (1-a)^2 + (8-6a)^2 + (4-10a)^2.$$

$$b_v = \frac{-(34a-26)}{2\cdot 3} = \frac{-2(17a-13)}{2\cdot 3} = \frac{13-17a}{3} = \frac{13}{3} - \frac{17a}{3}.$$

DEDUÇÃO USANDO FUNÇÕES QUADRÁTICAS

$$b_v = \frac{13}{3} - \frac{17a}{3}$$

$$d_1^2 + d_2^2 + d_3^2$$

$$\|$$

$$(1-(1a+b))^2 + (8-(6a+b))^2 + (4-(10a+b))^2$$

$$\|$$

$$\left[(1-a) - \left(\frac{13}{3} - \frac{17}{3}a\right)\right]^2 + \left[(8-6a) - \left(\frac{13}{3} - \frac{17}{3}a\right)\right]^2 + \left[(4-10a) - \left(\frac{13}{3} - \frac{17}{3}a\right)\right]^2$$

$$d_1^2 + d_2^2 + d_3^2$$

$$\|$$

$$\left[(1-a) - \left(\frac{13}{3} - \frac{17}{3}a\right)\right]^2 + \left[(8-6a) - \left(\frac{13}{3} - \frac{17}{3}a\right)\right]^2 + \left[(4-10a) - \left(\frac{13}{3} - \frac{17}{3}a\right)\right]^2$$

$$\|$$

$$\left[\left(1-\frac{13}{3}\right) - \left(1-\frac{17}{3}\right)a\right]^2 + \left[\left(8-\frac{13}{3}\right) - \left(6-\frac{17}{3}\right)a\right]^2 + \left[\left(4-\frac{13}{3}\right) - \left(10-\frac{17}{3}\right)a\right]^2$$

$$\|$$

$$\left[-\frac{10}{3} + \frac{14}{3}a\right]^2 + \left[\frac{11}{3} - \frac{1}{3}a\right]^2 + \left[-\frac{1}{3} - \frac{13}{3}a\right]^2$$

$$\|$$

$$\left[\frac{100}{9} - \frac{280}{9}a + \frac{196}{9}a^2\right] + \left[\frac{121}{9} - \frac{22}{9}a + \frac{1}{9}a^2\right] + \left[\frac{1}{9} + \frac{26}{9}a + \frac{169}{9}a^2\right]$$

$$\|$$

$$\frac{122}{9}a^2 - \frac{92}{3}a + \frac{74}{3}$$

DEDUÇÃO USANDO FUNÇÕES QUADRÁTICAS

$$d_1^2 + d_2^2 + d_3^2$$

$$\| $$

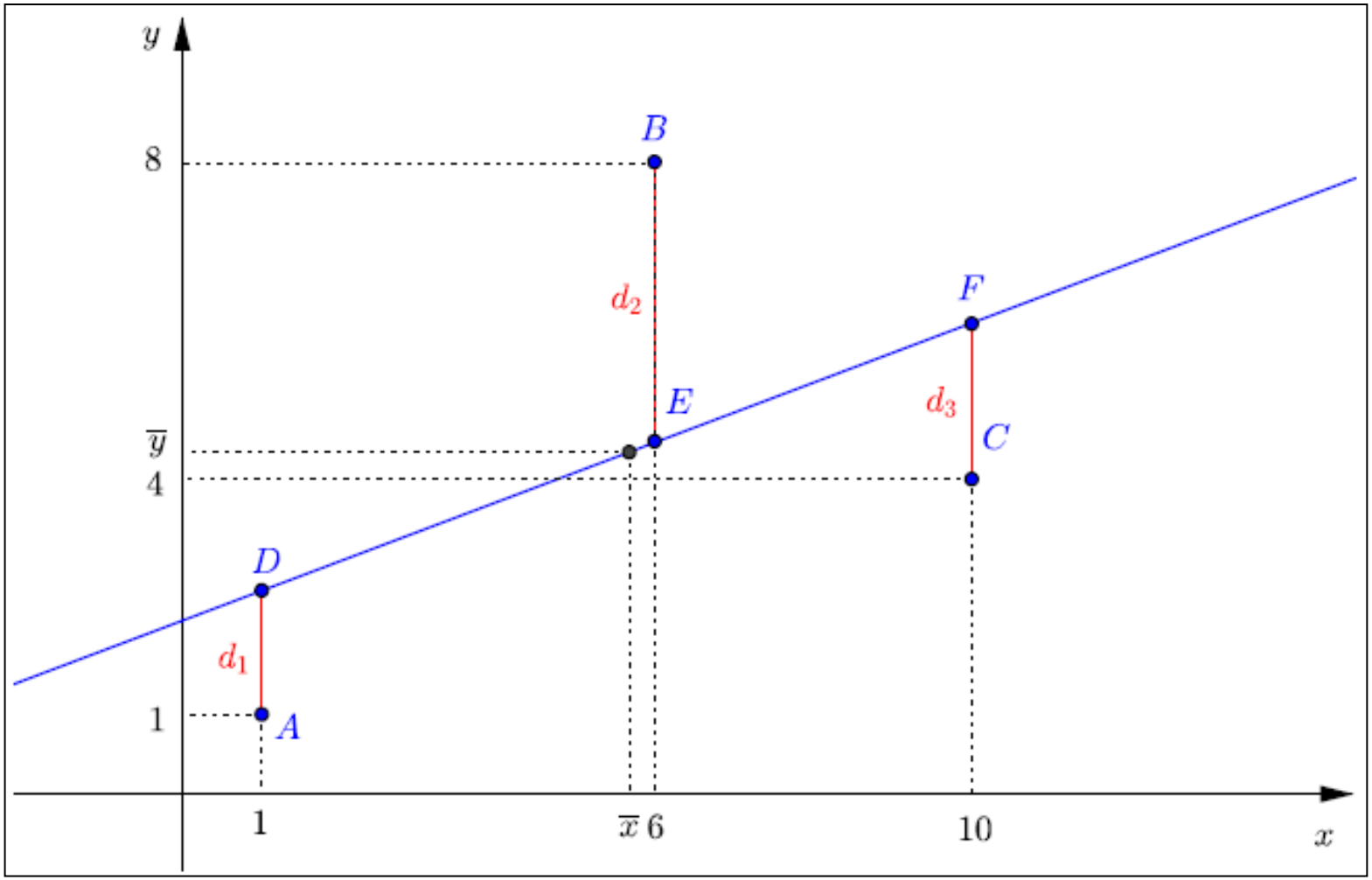
$$\frac{122}{9}a^2 - \frac{92}{3}a + \frac{74}{3}$$

$$a_v = -\frac{-\frac{92}{3}}{2\frac{122}{3}} = \frac{23}{61} = 0{,}377\ldots$$

$$b = \frac{13}{3} - a\frac{17}{3} = \frac{13}{3} - \frac{23}{61}\cdot\frac{17}{3} = \frac{134}{61} = 2{,}197\ldots$$

$$y = \frac{23}{61}a + \frac{134}{61}$$

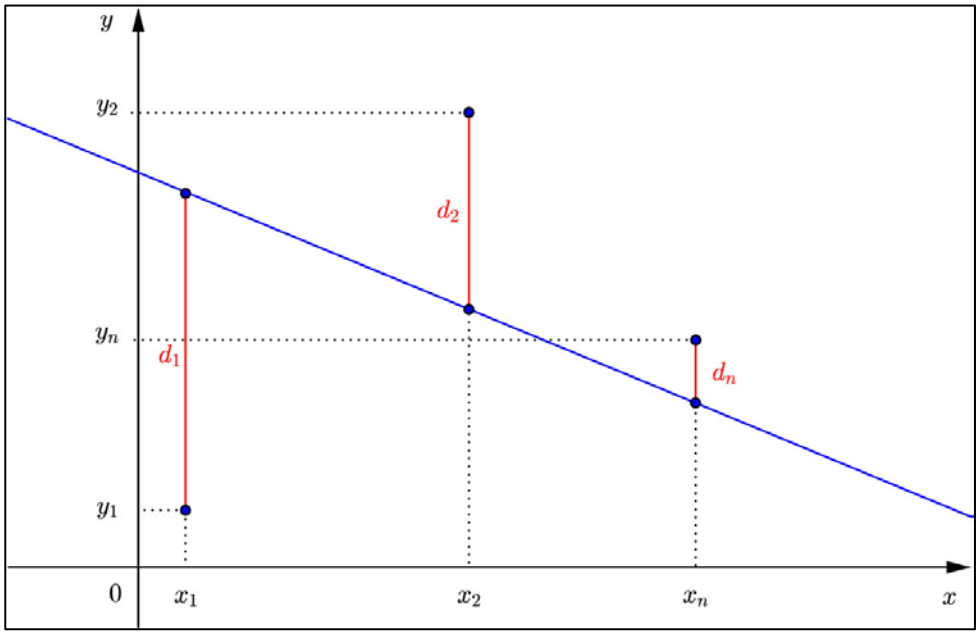
DEDUÇÃO USANDO FUNÇÕES QUADRÁTICAS



DEDUÇÃO USANDO FUNÇÕES QUADRÁTICAS

Sejam (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) pontos no plano cartesiano, nem todos com a mesma abscissa. Queremos minimizar em a e b (os coeficientes da **reta** $y = a\,x + b$):

$$\begin{aligned} d_1^2 + d_2^2 + \cdots + d_n^2 &= (y_1 - (ax_1 + b))^2 + (y_2 - (ax_2 + b))^2 + \cdots + (y_n - (ax_n + b))^2 \\ &= ((y_1 - ax_1) - b)^2 + ((y_2 - ax_2) - b)^2 + \cdots + ((y_n - ax_n) - b)^2. \end{aligned}$$



DEDUÇÃO USANDO FUNÇÕES QUADRÁTICAS

Para cada valor de a (isto é, considerando-se a como um parâmetro),

$$f(b)$$

$$\|$$

$$d_1^2 + d_2^2 + \cdots + d_n^2$$

$$\|$$

$$(y_1 - ax_1)^2 - 2(y_1 - ax_1)b + b^2 +$$
$$(y_2 - ax_2)^2 - 2(y_2 - ax_2)b + b^2 + \cdots +$$
$$(y_n - ax_n)^2 - 2(y_n - ax_n)b + b^2$$

$$\|$$

$$nb^2 -$$

$$2[(y_1 - ax_1) + (y_2 - ax_2) + \cdots + (y_n - ax_n)]b +$$
$$(y_1 - ax_1)^2 + (y_2 - ax_2)^2 + \cdots + (y_n - ax_n)^2.$$

DEDUÇÃO USANDO FUNÇÕES QUADRÁTICAS

$$
\begin{aligned}
b_v &= \frac{-[-2((y_1 - ax_1) + (y_2 - ax_2) + \cdots + (y_n - ax_n))]}{2 \cdot n} \\
&= \frac{(y_1 - ax_1) + (y_2 - ax_2) + \cdots + (y_n - ax_n)}{n} = \frac{y_1 + y_2 + \cdots + y_n}{n} - a\frac{x_1 + x_2 + \cdots + x_n}{n} \\
&= \bar{y} - a\bar{x}.
\end{aligned}
$$

$$
d_1^2 + d_2^2 + \cdots + d_n^2
$$

$$
\Vert
$$

$$
[(y_1 - ax_1) - (\bar{y} - a\bar{x})]^2 + [(y_2 - ax_2) - (\bar{y} - a\bar{x})]^2 + \cdots + [(y_n - ax_n) - (\bar{y} - a\bar{x})]^2
$$

$$
\Vert
$$

$$
[(y_1 - \bar{y}) - (x_1 - \bar{x})\,a]^2 + [(y_2 - \bar{y}) - (x_2 - \bar{x})\,a]^2 + \cdots + [(y_n - \bar{y}) - (x_n - \bar{x})\,a]^2
$$

DEDUÇÃO USANDO FUNÇÕES QUADRÁTICAS

$$d_1^2 + d_2^2 + \cdots + d_n^2$$

$$\|$$

$$(y_1 - \overline{y})^2 - 2a(y_1 - \overline{y})(x_1 - \overline{x}) + (x_1 - \overline{x})^2 a^2 +$$
$$(y_2 - \overline{y})^2 - 2a(y_2 - \overline{y})(x_2 - \overline{x}) + (x_2 - \overline{x})^2 a^2 + \cdots +$$
$$(y_n - \overline{y})^2 - 2a(y_n - \overline{y})(x_n - \overline{x}) + (x_n - \overline{x})^2 a^2 +$$

$$\|$$

$$\left[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\right] a^2$$
$$-2\left[(y_1 - \overline{y})(x_1 - \overline{x}) + (y_2 - \overline{y})(x_2 - \overline{x}) + \cdots + (y_n - \overline{y})(x_n - \overline{x})\right] a$$
$$+ (y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \cdots + (y_n - \overline{y})^2 .$$

DEDUÇÃO USANDO FUNÇÕES QUADRÁTICAS

$$a = \frac{(y_1-\overline{y})(x_1-\overline{x})+(y_2-\overline{y})(x_2-\overline{x})+\cdots+(y_n-\overline{y})(x_n-\overline{x})}{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2} \text{ e}$$

$$b = \overline{y}-\frac{(y_1-\overline{y})(x_1-\overline{x})+(y_2-\overline{y})(x_2-\overline{x})+\cdots+(y_n-\overline{y})(x_n-\overline{x})}{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}\overline{x}.$$

$$a=\frac{\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \quad \text{e} \quad b=\overline{y}-\frac{\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\overline{x}.$$

$$y=ax+b=ax+(\overline{y}-a\overline{x})=\overline{y}+a(x-\overline{x}).$$

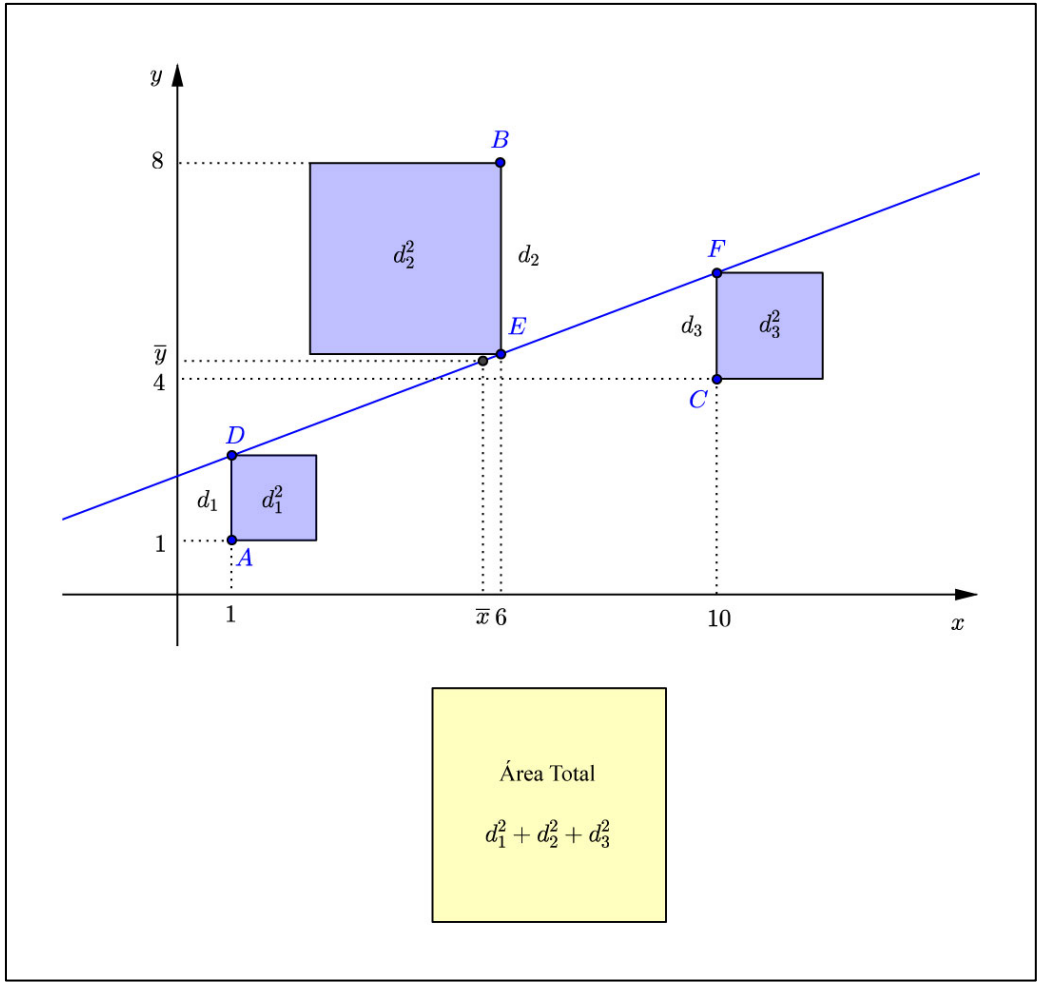
DEDUÇÃO USANDO FUNÇÕES QUADRÁTICAS

$$a = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{e} \quad b = \overline{y} - \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\overline{x}.$$

$$a = \frac{n\sum_{i=1}^{n}x_i y_i - \left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2} \quad \text{e} \quad b = \frac{\left(\sum_{i=1}^{n}x_i^2\right)\left(\sum_{i=1}^{n}y_i\right) - \left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}$$

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)}{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2}$$

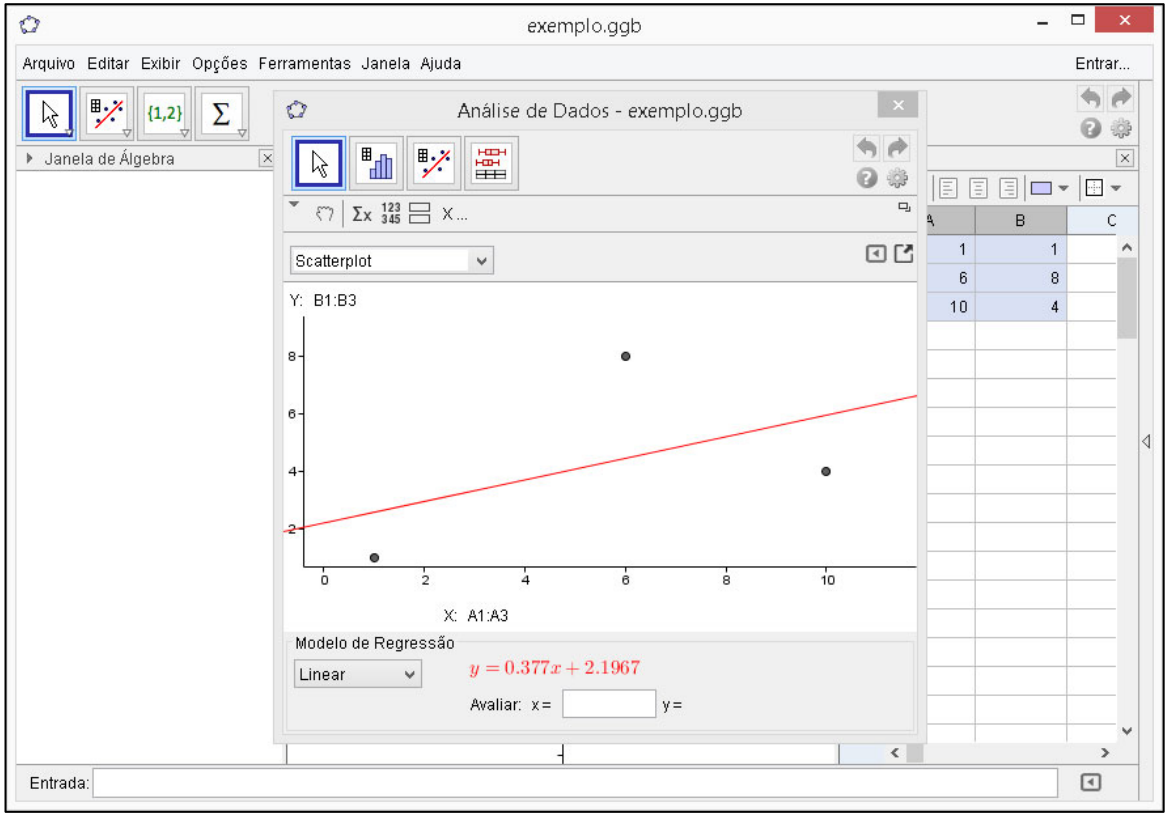
INTERPRETAÇÃO GEOMÉTRICA



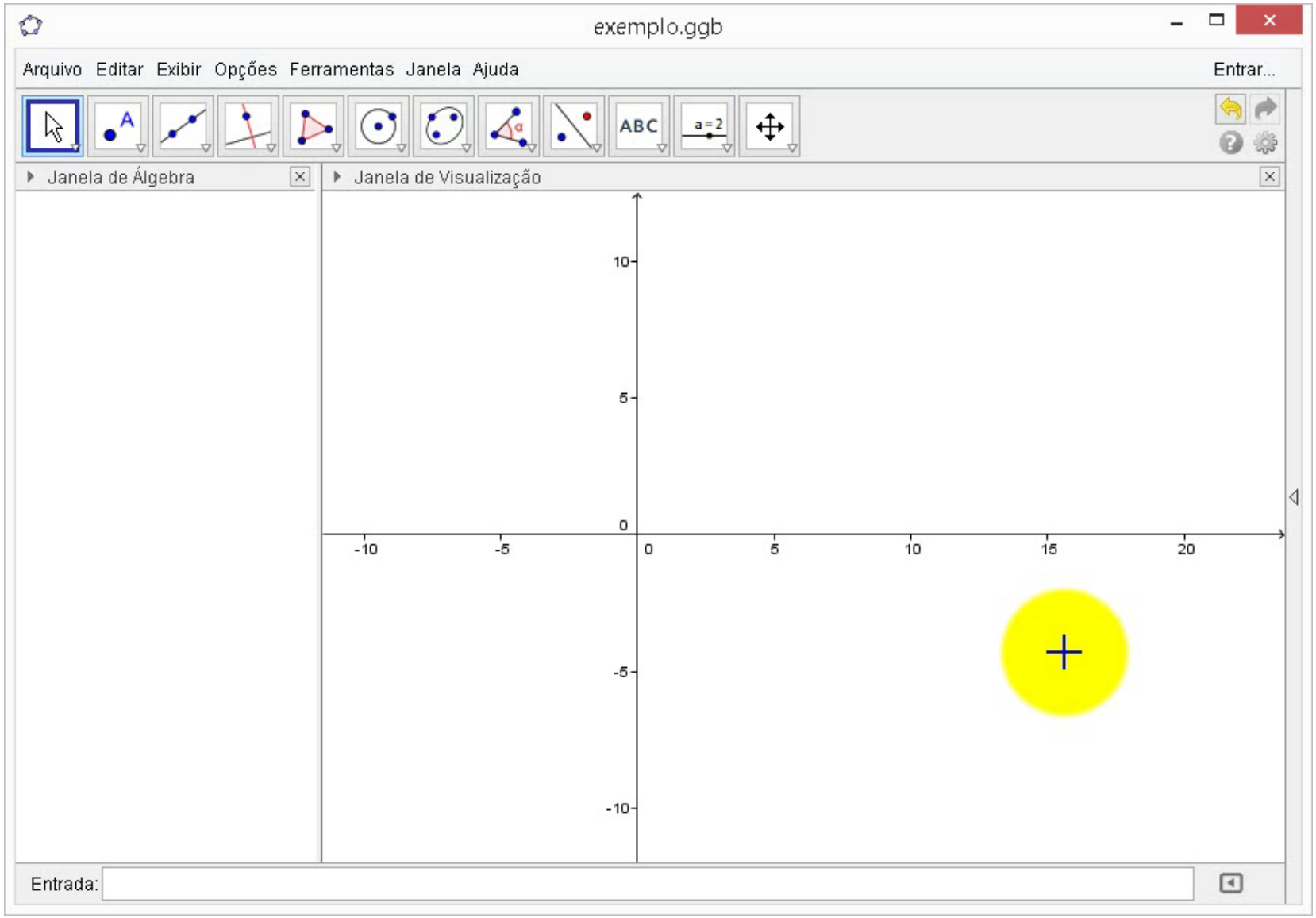
< **http://www.geogebratube.org/student/m84213/** >

USANDO TECNOLOGIA

GeoGebra, LibreOffice e Excel



< **http://youtu.be/n_LpIRwPKUI/** >


exemplo.ggb
Arquivo Editar Exibir Opções Ferramentas Janela Ajuda
Entrar...
ABC
a = 2
Janela de Álgebra
Janela de Visualização
10
5
0
-5
-10
-10
-5
0
5
10
15
20
Entrada:

APLICAÇÃO: A LEI DE ZIPF

(A)

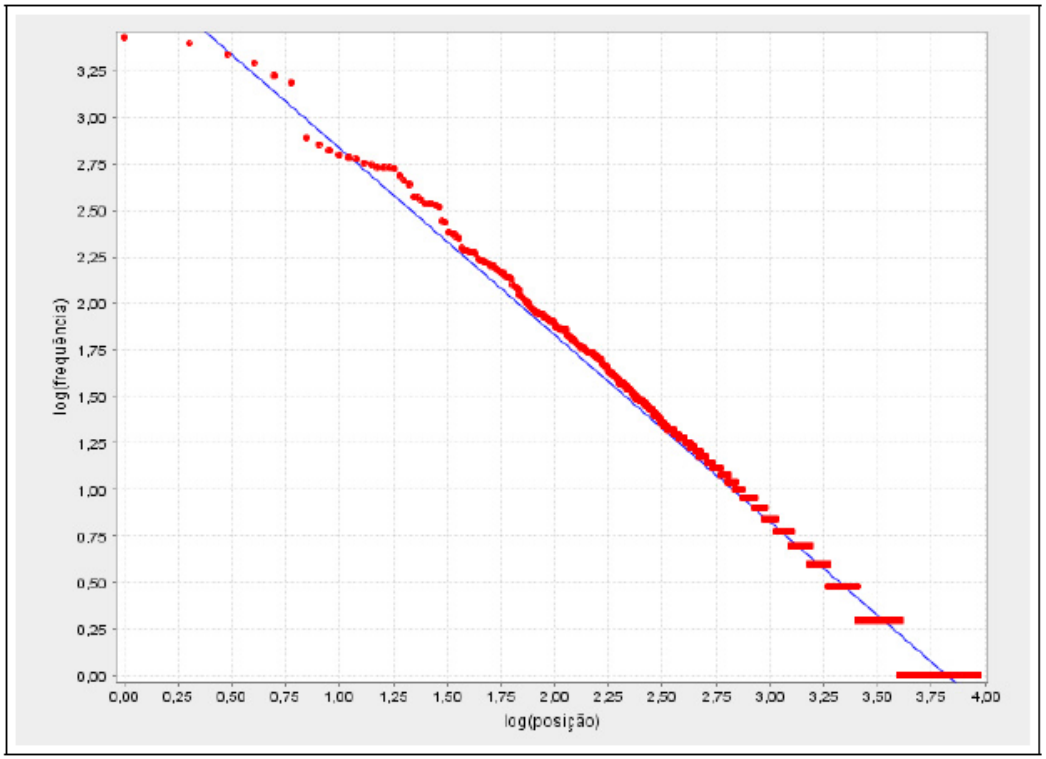
Posição (x)	Frequência (y)	Palavra
1	2684	que
2	2490	a
3	2186	e
4	1970	de
5	1671	o
6	1531	não
⋮	⋮	⋮
26	341	Capitu
⋮	⋮	⋮
141	56	Bentinho
⋮	⋮	⋮
9262	1	zanguei
9263	1	zás
9264	1	zeloso

(B)

$\tilde{x} = \log(x)$	$\tilde{y} = \log(y)$	Palavra
0,00000...	3,42878...	que
0,30103...	3,39619...	a
0,47712...	3,33965...	e
0,60205...	3,29446...	de
0,69897...	3,22297...	o
0,77815...	3,18497...	não
⋮	⋮	⋮
1,41497...	2,53275...	Capitu
⋮	⋮	⋮
2,14921...	1,74818...	Bentinho
⋮	⋮	⋮
3,96670...	0,00000...	zanguei
3,96675...	0,00000...	zás
3,96679...	0,00000...	zeloso

Tabela 3.1: Frequência das palavras em "Dom Casmurro" de Machado de Assis.

APLICAÇÃO: A LEI DE ZIPF



$$\tilde{y} = 3{,}837 - 1{,}005\tilde{x}$$

$$y = 6870{,}684x^{-1{,}005}$$

$$y = Cx^{-\alpha}$$

APLICAÇÃO: A LEI DE ZIPF

(A)

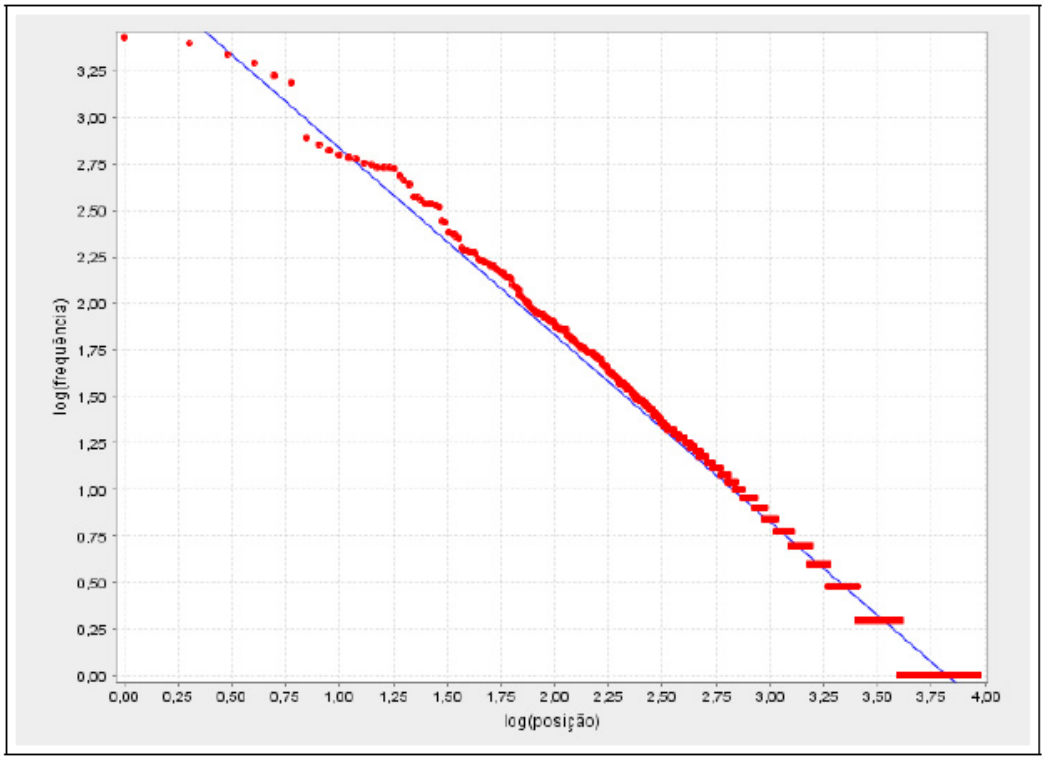
Posição (x)	Frequência (y)	Palavra
1	708	and
2	688	the
3	586	I
4	540	to
5	464	a
6	396	of
⋮	⋮	⋮
11	296	Romeo
⋮	⋮	⋮
22	178	Juliet
⋮	⋮	⋮
3781	1	yoke
3782	1	yon
3783	1	youngest

(B)

$\widetilde{x} = \log(x)$	$\widetilde{y} = \log(y)$	Palavra
0,00000...	2,85003...	and
0,30103...	2,83758...	the
0,47712...	2,76789...	I
0,60205...	2,73239...	to
0,69897...	2,66651...	a
0,77815...	2,59769...	of
⋮	⋮	⋮
1,04139...	2,47129...	Romeo
⋮	⋮	⋮
1,34242...	2,25042...	Juliet
⋮	⋮	⋮
3,57760...	0,00000...	yoke
3,57772...	0,00000...	yon
3,57783...	0,00000...	youngest

Tabela 3.2: Frequência das palavras em "Romeo and Juliet" de William Shakespeare.

APLICAÇÃO: A LEI DE ZIPF



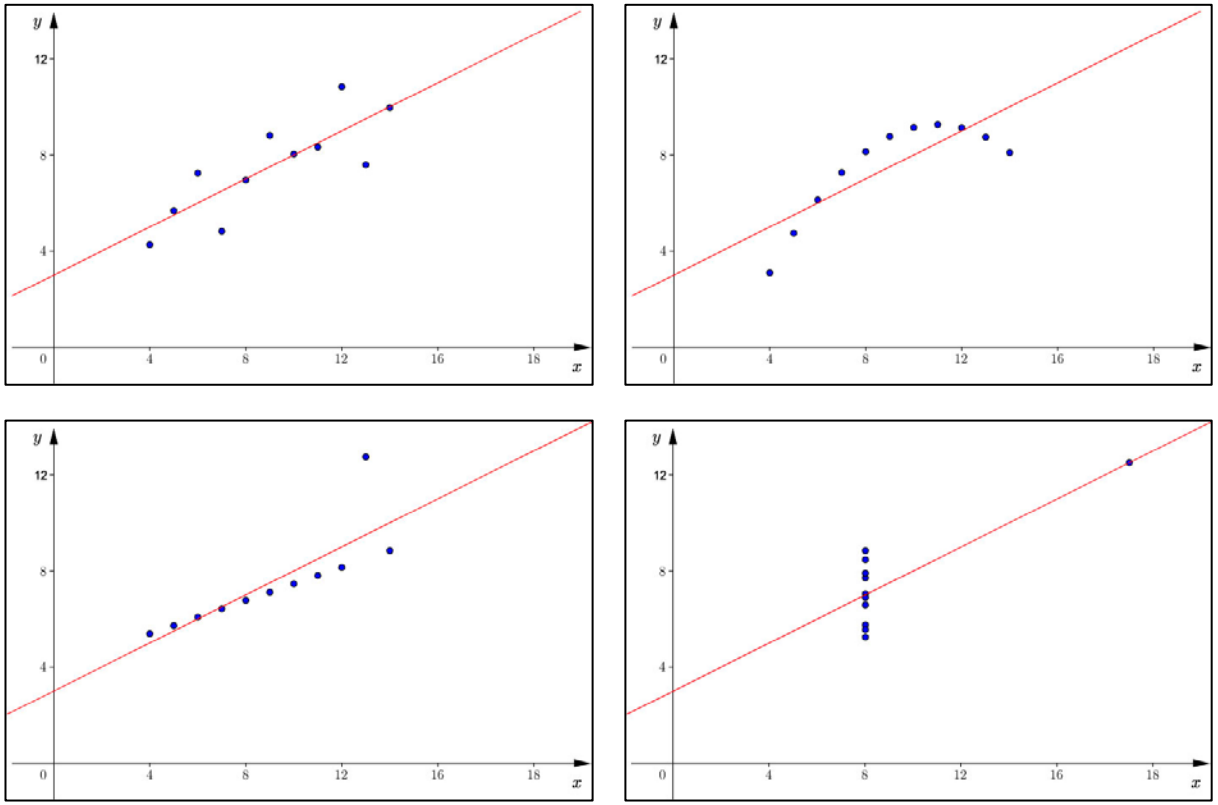
$$\tilde{y} = 3{,}674 - 1{,}070\tilde{x}$$

$$y = 4726{,}348x^{-1{,}070}$$

$$y = Cx^{-\alpha}$$

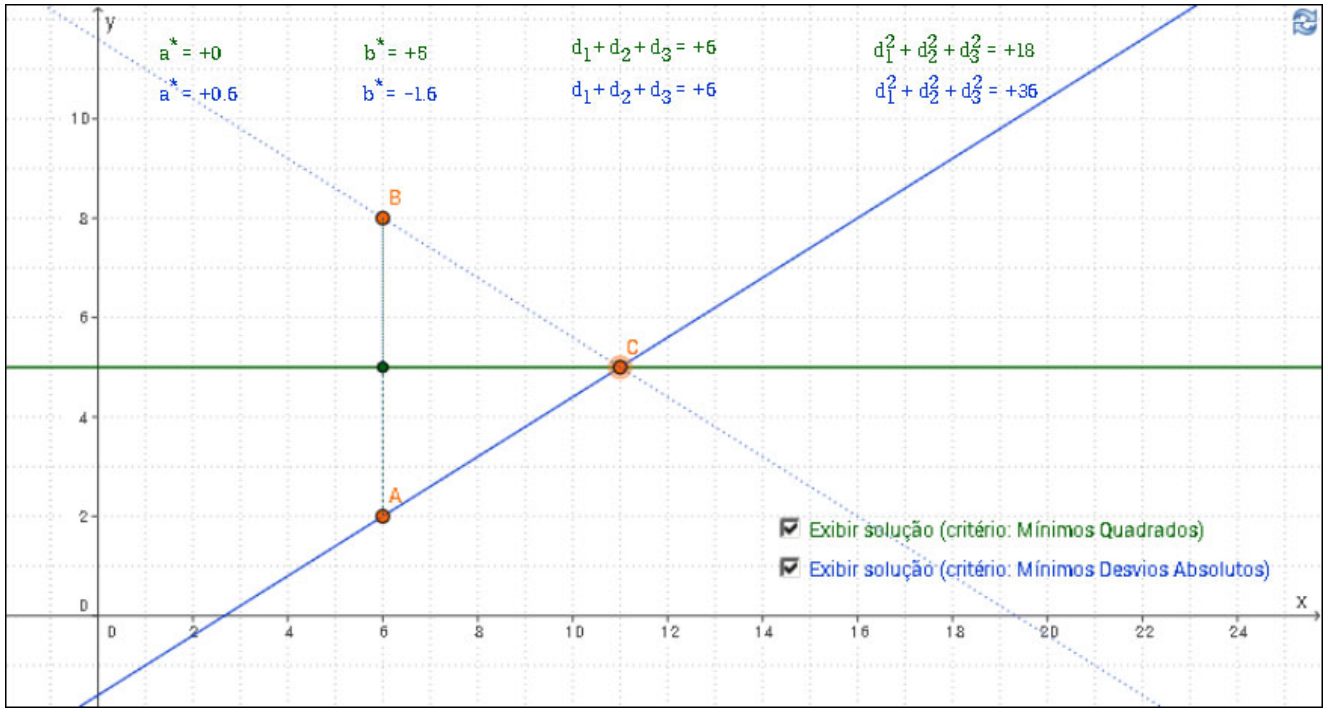
CONSIDERAÇÕES FINAIS

Como outros métodos e conceitos de Estatística, regressão é uma técnica que resume dados e, sendo assim, nem todos os aspectos do conjunto de dados são capturados.



CONSIDERAÇÕES FINAIS

O critério dos mínimos quadrados não é o único disponível para se encontrar uma reta de regressão. Temos, por exemplo, o método dos mínimos desvios absolutos.



< **http://www.geogebratube.org/student/m87161/** >

CONSIDERAÇÕES FINAIS

Recomendações curriculares de outros países indicam explicitamente a inclusão de correlação e regressão no Ensino Médio (nos Estados Unidos: *Principles and Standards* do NCTM e o *Common Core*).

Como sugestão de trabalho futuro, colocamos a questão de investigar como se realiza, nestes países, a inclusão destes temas no contexto escolar: livros didáticos, exemplos sugeridos, dificuldades encontradas, etc.

A referência *Teaching Statistics in School Mathematics – Challenges for Teaching and Teacher Education* de Batanero, Burrill & Reading nos parece ser um bom lugar para começar.

Obrigado!

hjbortol@vm.uff.br

Conteúdos Digitais

para o ensino e aprendizagem de matemática e estatística

http://www.uff.br/cdme/

GeoGebra

Instituto GeoGebra no Rio de Janeiro

http://www.uff.br/geogebra/