Introdução à Ciência dos Dados

Aula 04 – Inferência Estatística

Sumário

- Introdução
- Teste de hipótese
- Intervalo de confiança
- valor-p

Inferência Estatística

- Imagine um estudante que tirou as seguintes notas nas provas de Programação:

  30%; 23%; 40%; 30%; 98%

- O que podemos **concluir**?

Inferência Estatística

- Imagine um estudante que tirou as seguintes notas nas provas de Programação:

  30%; 23%; 40%; 30%; 98%

- O que podemos **concluir**?
  - Nada. *A estatística não pode provar nada com certeza.*
- Mas podemos usar inferência para tentar determinar a explicação mais provável para algum resultado

Inferência Estatística

- Imagine um estudante que tirou as seguintes notas nas provas de Programação:

  30%; 23%; 40%; 30%; 98%

- O que podemos **inferir**?

Inferência Estatística

- Imagine um estudante que tirou as seguintes notas nas provas de Programação:

  30%; 23%; 40%; 30%; 98%

- O que podemos **inferir**?
  - Com base em anos lecionando a disciplina de Programação, o professor tem dados suficientes para inferir que este estudante provavelmente colou na última prova.
  - O professor pode ter certeza disso? **Não.** Mas ele tem uma boa chance de estar certo....

Inferência Estatística

- Imagine que um estranho chegue em sua cidade e te ofereça a seguinte aposta:
  - Ele ganha R$1.000 se tirar 6 em um lance de um dado
  - Você ganha R$500 se der qualquer outro número

Inferência Estatística

- Imagine que um estranho chegue em sua cidade e te ofereça a seguinte aposta:
  - Ele ganha R$1.000 se tirar 6 em um lance de um dado
  - Você ganha R$500 se der qualquer outro número
- Ele então joga e tira dez vezes seguidas o 6, levando R$10.000

Inferência Estatística

- Imagine que um estranho chegue em sua cidade e te ofereça a seguinte aposta:
  - Ele ganha R$1.000 se tirar 6 em um lance de um dado
  - Você ganha R$500 se der qualquer outro número
- Ele então joga e tira dez vezes seguidas o 6, levando R$10.000
  - Uma explicação possível é que ele teve muita sorte
  - Uma explicação alternativa é que o dado é viciado, pois a probabilidade de tirar o 6 dez vezes seguidas é de uma em 60 milhões
- Você não pode **provar** que ele trapaceou, mas no mínimo deve ficar desconfiado

Inferência Estatística

- Quando um evento muito raro acontece, podemos desconfiar
- Mas cuidado: coisas extremamente raras acontecem
  - Linda Cooper foi atingida 4 vezes por raios
  - O estudante que tirou 98% na última prova pode ter resolvido se dedicar à Programação depois que já sabia que havia sido reprovado em Cálculo
- Mas um padrão improvável deve ser investigado
  - É assim que Comissão de Valores Imobiliários pega operações com informação privilegiada

Inferência Estatística

- Será que um novo remédio é efetivo no tratamento de doenças cardíacas?
- Será que celulares provocam câncer?
- Será que o meu protocolo é mais eficiente que o já existente?
- Será que a prefeitura da minha cidade está gastando o dinheiro honestamente?
- Será que o remédio Paracetamol cura COVID-19?

Inferência Estatística

- Será que um novo remédio é efetivo no tratamento de doenças cardíacas?
- Será que celulares provocam câncer?
- Será que o meu protocolo é mais eficiente que o já existente?
- Será que a prefeitura da minha cidade está gastando o dinheiro honestamente?
- Será que o remédio Paracetamol cura COVID-19?

A inferência não responde a essas perguntas, mas nos diz o que é provável e o que é improvável.

Inferência Estatística

- Suponha que 91 em cada 100 pacientes recebendo uma nova medicação mostrem uma melhora acentuada, em comparação com 49 em 100 do grupo de controle.
  - Ainda é possível que esse resultado impressionante não esteja relacionado com a nova droga. Mas essa explicação é bem menos provável....

Inferência Estatística

- Suponha que 91 em cada 100 pacientes recebendo uma nova medicação mostrem uma melhora acentuada, em comparação com 49 em 100 do grupo de controle.

  1) se o medicamento não tem efeito, raramente veríamos uma variação de resultados dessa dimensão entre os que receberam e os que não receberam o tratamento

  2) portanto, é muito improvável que o medicamento não tenha efeito positivo

  3) a explicação alternativa, e mais provável, é que o medicamente tenha efeito positivo

Inferência Estatística

- Inferência estatística é o processo pelo qual os dados falam conosco, possibilitando-nos tirar conclusões significativas
  - Dados e probabilidade, com ajuda do teorema central do limite
  - Aceitar ou rejeitar explicações/hipóteses com base na sua relativa probabilidade

Inferência Estatística

- Teste de hipóteses
  - *Hipótese Nula* implícita ou explícita
    - Premissa de partida, que será rejeitada ou não
    - Se rejeitada, aceitamos alguma hipótese alternativa que seja mais consistente com os dados observados
    - Não será provada verdadeira; apenas pode-se falhar em rejeitá-la
  - *Hipótese Alternativa*
    - Conclusão que precisa ser verdadeira se é para rejeitar a hipótese nula

Inferência Estatística

Exemplo: Droga para prevenir malária

- *Hipótese Nula*: nova droga não é mais efetiva em prevenir a malária do que um placebo
- *Hipótese Alternativa*: nova droga pode ajudar a prevenir a malária
- *Dados:* um grupo aleatório recebe a nova droga e um grupo de controle recebe placebo
- *Resultado:* grupo que recebe a droga tem muito menos casos de malária que o grupo de controle

Inferência Estatística

Exemplo: Droga para prevenir malária

- *Hipótese Nula*: nova droga não é mais efetiva em prevenir a malária do que um placebo
- *Hipótese Alternativa*: nova droga pode ajudar a prevenir a malária
- *Dados:* um grupo aleatório recebe a nova droga e um grupo de controle recebe placebo
- *Resultado:* grupo que recebe a droga tem muito menos casos de malária que o grupo de controle

Resultado extremamente improvável se a droga não tivesse impacto medicinal. Por isso, rejeitamos a hipótese nula e aceitamos a alternativa. Ou seja, essa nova droga pode ajudar a prevenir malária.

Inferência Estatística

Exemplo: Reincidência em crimes

- *Hipótese Nula*: tratamento para abuso de substâncias químicas para detentos não reduz sua taxa de reincidência
- *Hipótese Alternativa*: tratamento para abuso de substâncias químicas para detentos reduzirá a probabilidade de reincidência
- *Dados:* um grupo aleatório de detentos recebe tratamento para abuso de substâncias e o grupo de controle não recebe
- *Resultado:* Após 5 anos, ambos os grupos têm índices similares de reincidência

Inferência Estatística

Exemplo: Reincidência em crimes

- *Hipótese Nula*: tratamento para abuso de substâncias químicas para detentos não reduz sua taxa de reincidência
- *Hipótese Alternativa*: tratamento para abuso de substâncias químicas para detentos reduzirá a probabilidade de reincidência
- *Dados:* um grupo aleatório de detentos recebe tratamento para abuso de substâncias e o grupo de controle não recebe
- *Resultado:* Após 5 anos, ambos os grupos têm índices similares de reincidência

Não podemos rejeitar a hipótese nula. Os dados não nos deram razão para descartar nossa premissa inicial de que o tratamento para abuso de substâncias químicas não é uma ferramenta efetiva para diminuir a reincidência.

Inferência Estatística

- Muitas vezes, a hipótese nula é criada com a esperança de que seja rejeitada. Nesses exemplos, o “sucesso” da pesquisa envolvia rejeitar a hipótese nula.
- A pergunta a ser respondida quantitativamente:
  - Se a hipótese nula for **verdadeira**, qual é a probabilidade de observar esse padrão de dados por puro acaso?

Inferência Estatística

- Mas o quanto a hipótese nula deve ser implausível para podermos rejeitá-la e recorrer a alguma explicação alternativa?
  - 5% (0,05) é um dos limiares mais comuns
  - Chamado de nível de significância
  - Limite superior para a probabilidade de observação de algum padrão de dados se a hipótese nula fosse verdadeira
  - Podemos rejeitar a hipótese nula no nível 0,05 se a chance de obter um resultado no mínimo tão extremo quanto o que observamos se a hipótese nula for verdadeira for **menor que 5%**

Inferência Estatística

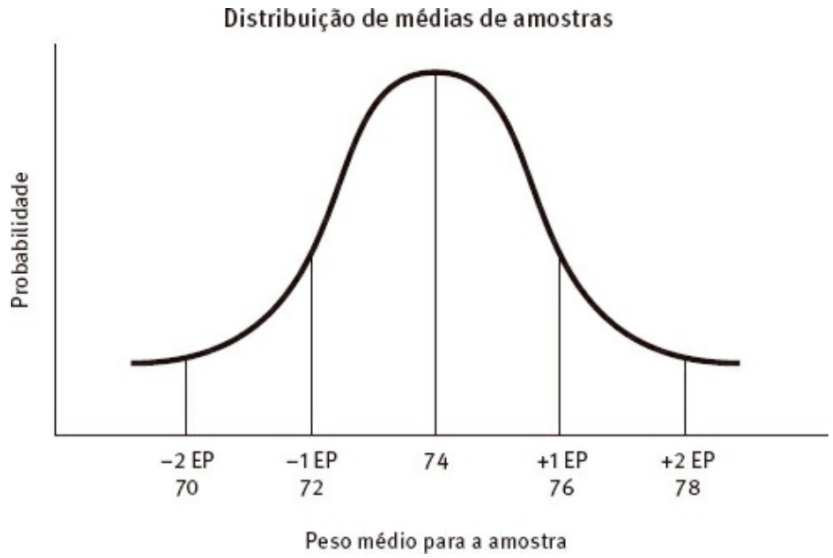
- Exemplo do ônibus perdido
  - Pessoas no ônibus (amostra): 60
  - Peso médio da população: 74 KG
  - Desvio padrão da população: 16 KG
  - Erro padrão: 16/sqrt(60) = 2,1

Inferência Estatística

- Exemplo do ônibus perdido
  - Pessoas no ônibus (amostra): 60
  - Peso médio da população: 74 KG
  - Desvio padrão da população: 16 KG
  - Erro padrão: 16/sqrt(60) = 2,1

Espera-se que 95% de todas as amostras de 60 pessoas tenham peso médio dentro de dois erros padrões em relação à média da população, ou aproximadamente entre 70 e 78 quilos.

Inversamente, apenas 5 vezes em 100 uma amostra de 60 pessoas teria um peso médio menor que 70 ou maior que 78 quilos.

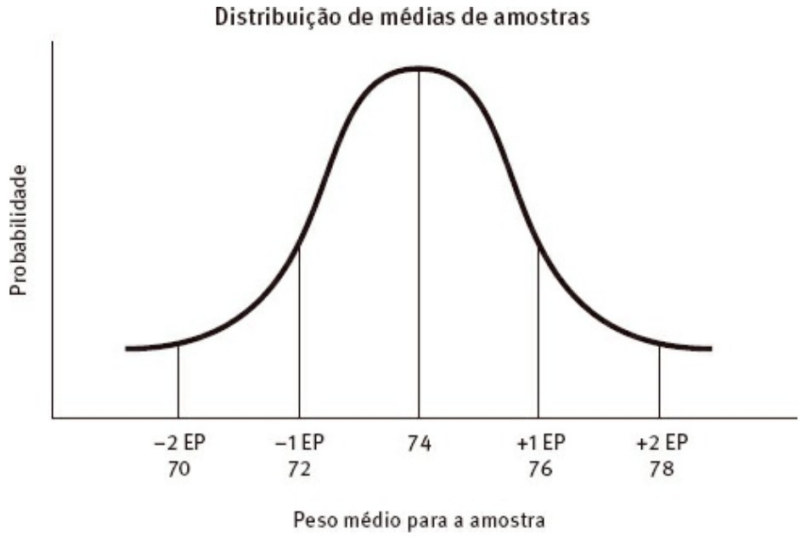


Inferência Estatística

- Exemplo do ônibus perdido
  - Pessoas no ônibus (amostra): 60
  - Peso médio da população: 74 KG
  - Desvio padrão da população: 16 KG
  - Erro padrão: 16/sqrt(60) = 2,1
  - Média de peso no ônibus: 86 KG

Espera-se que 95% de todas as amostras de 60 pessoas tenham peso médio dentro de dois erros padrões em relação à média da população, ou aproximadamente entre 70 e 78 quilos.

Inversamente, apenas 5 vezes em 100 uma amostra de 60 pessoas teria um peso médio menor que 70 ou maior que 78 quilos.



Inferência Estatística

- Exemplo do ônibus perdido
  - Pessoas no ônibus (amostra): 60
  - Peso médio da população: 74 KG
  - Desvio padrão da população: 16 KG
  - Erro padrão: 16/sqrt(60) = 2,1
  - Média de peso no ônibus: 86 KG

Como 86 KG é mais de 2 erros padrões acima da média, podemos rejeitar a hipótese nula de que o ônibus contém integrantes da população em análise. Ou seja:

1) o peso médio da amostra cai em uma faixa que deveria ocorrer somente 5 em cada 100 vezes, se a amostra viesse da população em análise;
2) podemos rejeitar a hipótese com um nível de significância 0,05;
3) em média, 95 vezes em 100 teremos rejeitado corretamente a hipótese nula, e apenas 5 vezes em 100 estaremos errados.

Inferência Estatística

- Mas qual a probabilidade específica de obter um resultado no mínimo tão extremo quanto o que você observou se a hipótese nula for verdadeira?
  - **valor-p** ou ***p-value***
  - No exemplo anterior, 86 KG corresponde a 5,7 erros padrões acima da média
  - A probabilidade da amostra ser da população em análise é de apenas **valor-p=0,0001**

Inferência Estatística

“Descoberta relação entre autismo e tamanho do cérebro”

- Exames de imagem em 59 crianças com autismo e 38 crianças sem autismo
- Resultados: crianças com autismo têm cérebros que são até 10% maiores que das crianças da mesma idade sem autismo

Inferência Estatística

“Descoberta relação entre autismo e tamanho do cérebro”

- Pergunta: essa descoberta pode ser válida, considerando que os estudos foram baseados em apenas 97 crianças ao todo?

Inferência Estatística

“Descoberta relação entre autismo e tamanho do cérebro”

- Pergunta: essa descoberta pode ser válida, considerando que os estudos foram baseados em apenas 97 crianças ao todo?
  - Sim, pelo teorema central do limite
- A probabilidade de observar as diferenças seria de 2 em 1000 (p = 0,002), se de fato não houvesse diferença real entre os grupos

Inferência Estatística

“Descoberta relação entre autismo e tamanho do cérebro”

- Hipótese nula: não há diferença nos cérebros de crianças com autismo e sem autismo
- Hipótese alternativa: cérebros de crianças com autismo são fundamentalmente diferentes

Inferência Estatística

“Descoberta relação entre autismo e tamanho do cérebro”

- Hipótese nula: não há diferença nos cérebros de crianças com autismo e sem autismo
- Hipótese alternativa: cérebros de crianças com autismo são fundamentalmente diferentes

Crianças com autismo: **média = 1310,4 cm^3**

Crianças sem autismo: **média = 1238,8 cm^3**

Diferença: 71,6 cm^3

Qual a chance de termos diferença, caso a hipótese nula seja verdadeira?

Inferência Estatística

“Descoberta relação entre autismo e tamanho do cérebro”

Crianças com autismo: **média = 1310,4 cm3 / EP = 13 cm3**

Crianças sem autismo: **média = 1238,8 cm3 / EP = 18 cm3**

- 95 vezes em 100 (95% de confiança), o intervalo de 1310,4 +/- 26 conterá o valor médio para todas as crianças com autismo
- 95 vezes em 100 (95% de confiança), o intervalo de 1238,8 +/- 36 incluirá o valor médio para crianças na população sem autismo

Inferência Estatística

“Descoberta relação entre autismo e tamanho do cérebro”

Crianças com autismo: **média = 1310,4 cm3 / EP = 13 cm3**

Crianças sem autismo: **média = 1238,8 cm3 / EP = 18 cm3**

- 95 vezes em 100 (95% de confiança), o intervalo de 1310,4 +/- 26 conterá o valor médio para todas as crianças com autismo
- 95 vezes em 100 (95% de confiança), o intervalo de 1238,8 +/- 36 incluirá o valor médio para crianças na população sem autismo

Inferência Estatística

“Descoberta relação entre autismo e tamanho do cérebro”

Ok, os intervalos de confiança não se sobrepõem. Mas qual a probabilidade de observarmos esses valores se realmente não houver diferença no tamanho dos cérebros entre os dois grupos?

- Podemos calcular o **valor-p**

Inferência Estatística

“Descoberta relação entre autismo e tamanho do cérebro”

Como calcular o valor-p?

- Se pegarmos duas amostras grandes da mesma população, seria de esperar que tenham médias bastante similares, ou idênticas.
- Ou seja, a diferença entre as médias deve ser próxima de zero
- O teorema central do limite nos diz que a diferença entre duas médias estará distribuída aproximadamente como uma distribuição Normal
- Se duas amostras provêm da mesma população, então, em cerca de 68 casos em 100, a diferença entre as médias estará dentro de um erro padrão de zero. E cerca de 95 casos em 100, a diferença estará dentro de dois erros padrões. E em 99,7 casos em 100, a diferença estará dentro de três erros padrões....

Inferência Estatística

"Descoberta relação entre autismo e tamanho do cérebro"

- **Como calcular o valor-p?**
- Se pegarmos duas amostras grandes da mesma população, seria de esperar que tenham médias bastante similares, ou idênticas.
- Ou seja, a diferença entre as médias deve ser próxima de zero
- O teorema central do limite nos diz que a diferença entre duas médias estará distribuída aproximadamente como uma distribuição Normal
- Se duas amostras provêm da mesma população, então, em cerca de 68 casos em 100, a diferença entre as médias estará dentro de um erro padrão de zero. E cerca de 95 casos em 100, a diferença estará dentro de dois erros padrões. E em 99,7 casos em 100, a diferença estará dentro de três erros padrões....

Erro padrão da diferença: 22,7 (Veja como calcular no próximo slide)

Diferença entre médias: 71,6 → Mais que 3 erros padrões

Valor-p = 0,002

Inferência Estatística

Cálculo do erro padrão para uma diferença de médias

Fórmula para comparar duas médias:

$$\frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

$\longrightarrow$ o numerador fornece o valor da diferença entre as médias

$\longrightarrow$ o denominador fornece o erro padrão para uma diferença entre as médias das duas amostras

onde:

$\overline{x}$ = média da amostra x

$\overline{y}$ = média da amostra y

s_x = desvio padrão para a amostra x

s_y = desvio padrão para a amostra y

n_x = número de observações na amostra x

n_y = número de observações na amostra y

Inferência Estatística

X’ = 1310,4
Y’ = 1238,8

Nx = 59
Ny = 38

Sx = EPx * sqrt(Nx) = 13 * 7,68 = 99,84
Sy = EPy * sqrt(Ny) = 18 * 6,16 = 110,95

Erro padrão da diferença: 22,7

X’ - Y’ = 1310,4 – 1238,8 = 71,6

Razão: 71,6 / 22,7 = 3,15

Diferença entre as médias está 3,15 EPs do zero

Cálculo do erro padrão para uma diferença de médias

Fórmula para comparar duas médias:

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

→ o numerador fornece o valor da diferença entre as médias

→ o denominador fornece o erro padrão para uma diferença entre as médias das duas amostras

onde:

$\bar{x}$ = média da amostra x

$\bar{y}$ = média da amostra y

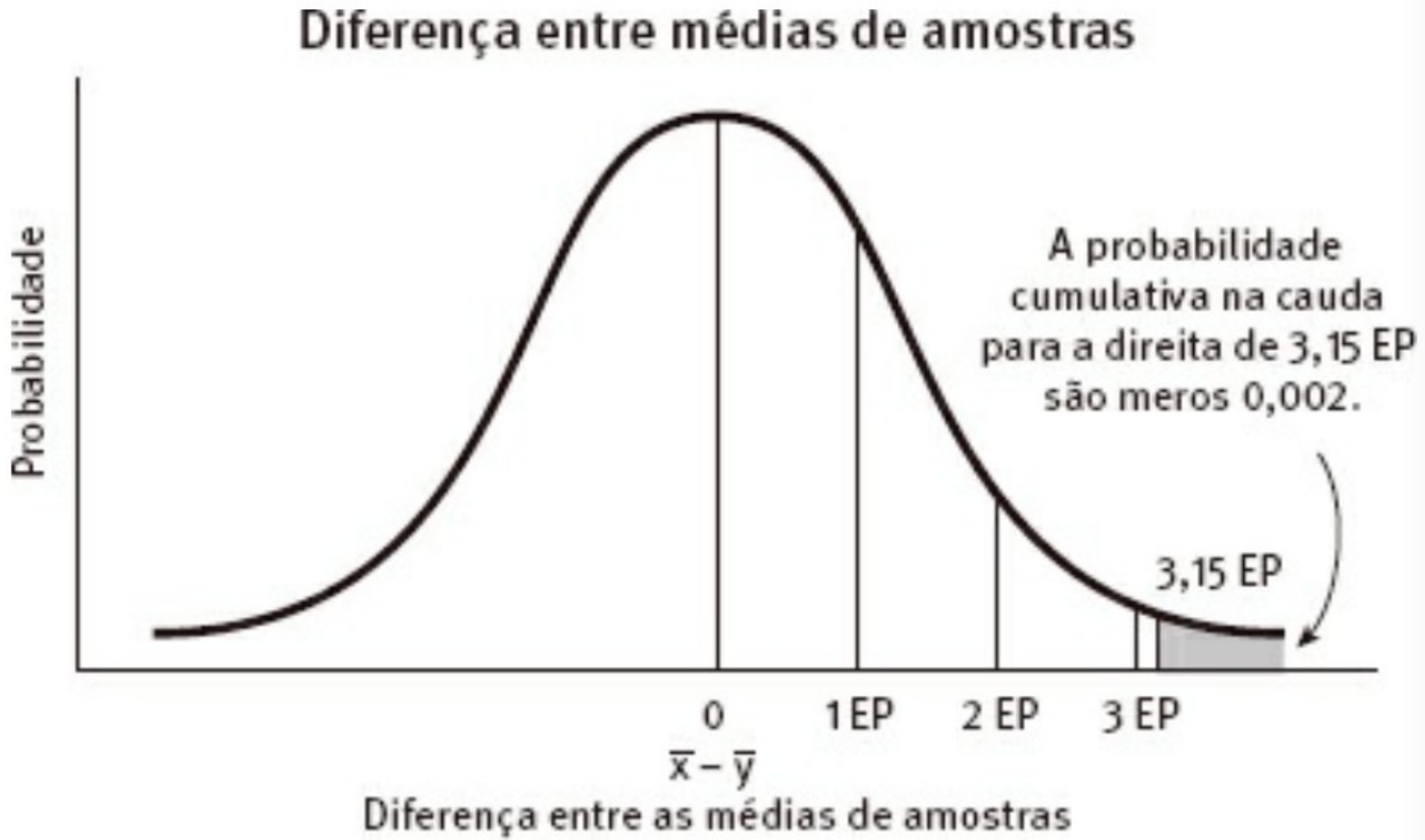
s_x = desvio padrão para a amostra x

s_y = desvio padrão para a amostra y

n_x = número de observações na amostra x

n_y = número de observações na amostra y

Inferência Estatística



Inferência Estatística

- Hipótese nula é rejeitada se probabilidade de encontrarmos os dados observados for menor que um limiar/nível de significância
- Qual o melhor limiar para rejeitar uma hipótese nula?
  - Depende das circunstâncias de cada caso
- Quanto menor o nível de significância, menos provável que a rejeição ocorra, e mais peso estatístico a rejeição tem
- Quanto maior o nível de significância, mais provável de rejeitar a hipótese nula

Inferência Estatística

- Qual o melhor limiar para rejeitar uma hipótese nula?
  - **Erro Tipo I**: rejeitar equivocadamente uma hipótese nula (Falso Positivo)
    - Quanto maior o limiar, maiores as chances
    - Ex. Inocentes presos, drogas que não funcionam no mercado
  - **Erro Tipo II:** não rejeitar uma hipótese nula que deveria ser rejeitada (Falso Negativo)
    - Quanto menor o limiar, maiores as chances
    - Ex. Culpados soltos, drogas que funcionam fora do mercado
- Qual tipo de erro é pior?

Inferência Estatística

- Qual o melhor limiar para rejeitar uma hipótese nula?
  - **Erro Tipo I**: rejeitar equivocadamente uma hipótese nula (Falso Positivo)
    - Quanto maior o limiar, maiores as chances
    - Ex. Inocentes presos, drogas que não funcionam no mercado
  - **Erro Tipo II:** não rejeitar uma hipótese nula que deveria ser rejeitada (Falso Negativo)
    - Quanto menor o limiar, maiores as chances
    - Ex. Culpados soltos, drogas que funcionam fora do mercado
- Qual tipo de erro é pior?
  - Depende das circunstâncias

Inferência Estatística

		Situação real	
		H_0 é verdadeira	H_0 é falsa
Nossa Decisão	Rejeitar H_0	***Erro Tipo I*** (Rejeitar H_0, quando H_0 é verdadeira)	Decisão correta
	Não Rejeitar H_0	Decisão correta	***Erro Tipo II*** (Não Rejeitar H_0, quando H_0 é falsa)

Inferência Estatística

- Qual o melhor limiar para rejeitar uma hipótese nula?

  1) Filtro de spam:

  - Hipótese nula: a mensagem **não** é spam

Inferência Estatística

- Qual o melhor limiar para rejeitar uma hipótese nula?

  1) Filtro de spam:

  - Hipótese nula: a mensagem **não** é spam
  - Erro Tipo I: mensagem que não é spam excluída
  - Erro Tipo II: spam continuar na caixa de entrada

  Erro Tipo II é mais aceitável. Quanto menor o limiar, menores as chances de Erro Tipo I.

Inferência Estatística

- Qual o melhor limiar para rejeitar uma hipótese nula?

  2) Diagnóstico de Câncer:

  - Hipótese nula: a pessoa **não** possui câncer

Inferência Estatística

- Qual o melhor limiar para rejeitar uma hipótese nula?

  2) Diagnóstico de Câncer:

  - Hipótese nula: a pessoa **não** possui câncer
  - Erro Tipo I: pessoa **sem** câncer é diagnosticada (falso positivo)
  - Erro Tipo II: pessoa **com** câncer não é diagnosticada (falso negativo)

  Erro Tipo I é mais aceitável. Quanto maior o limiar, menores as chances de Erro Tipo II.

Inferência Estatística

- Qual o melhor limiar para rejeitar uma hipótese nula?

  3) Captura de terroristas:

  - Hipótese nula: a pessoa **não** é terrorista

Inferência Estatística

- Qual o melhor limiar para rejeitar uma hipótese nula?

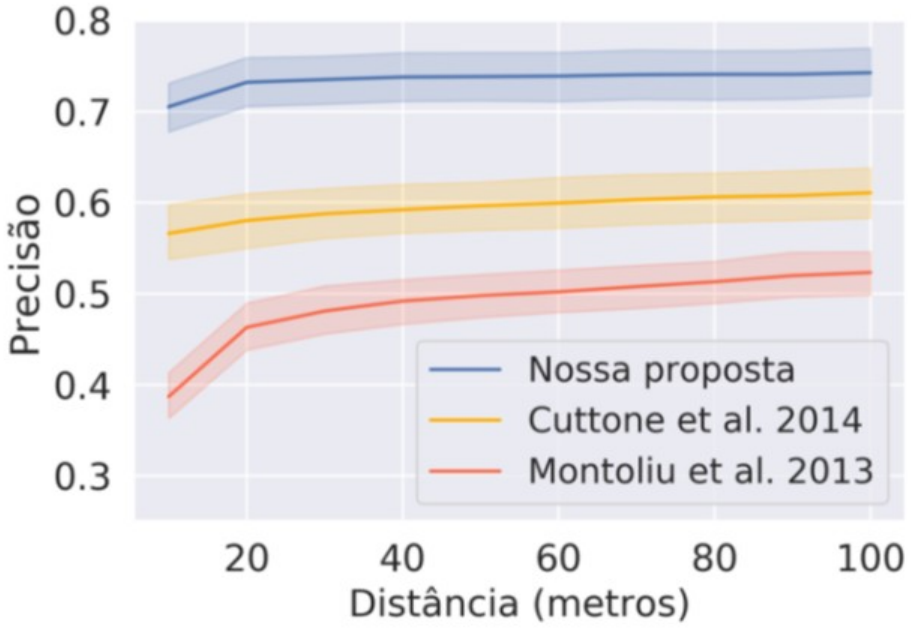
  3) Captura de terroristas:

  - Hipótese nula: a pessoa **não** é terrorista
  - Erro Tipo I: **não** terrorista preso e enviado para Guantánamo (falso positivo)
  - Erro Tipo II: terrorista liberado (falso negativo)

  Nenhum erro é aceitável. Porém, um único terrorista solto pode ser catastrófico. Decisão complicada.

Inferência Estatística

- Exemplos



30 simulações, comparação entre soluções com base nessa amostra de tamanho 30. Essa é uma estimativa da população.

Análise com 2000 usuários

Análise da Google com muitos milhares de usuários

Inferência Estatística

- Exemplos

Table 1. Baseline Characteristics of Patients Infected With 2019-nCoV

	No. (%)			
	Total (N = 138)	ICU (n = 36)	Non-ICU (n = 102)	*P* Value[a]
Signs and symptoms				
Fever	136 (98.6)	36 (100)	100 (98.0)	>.99
Fatigue	96 (69.6)	29 (80.6)	67 (65.7)	.10
Dry cough	82 (59.4)	21 (58.3)	61 (59.8)	.88
Anorexia	55 (39.9)	24 (66.7)	31 (30.4)	<.001
Myalgia	48 (34.8)	12 (33.3)	36 (35.3)	.83
Dyspnea	43 (31.2)	23 (63.9)	20 (19.6)	<.001
Expectoration	37 (26.8)	8 (22.2)	29 (28.4)	.35
Pharyngalgia	24 (17.4)	12 (33.3)	12 (11.8)	.003
Diarrhea	14 (10.1)	6 (16.7)	8 (7.8)	.20
Nausea	14 (10.1)	4 (11.1)	10 (9.8)	>.99
Dizziness	13 (9.4)	8 (22.2)	5 (4.9)	.007
Headache	9 (6.5)	3 (8.3)	6 (5.9)	.70
Vomiting	5 (3.6)	3 (8.3)	2 (2.0)	.13
Abdominal pain	3 (2.2)	3 (8.3)	0 (0)	.02

From one of the most cited COVID-19 papers
Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus Infected Pneumonia in Wuhan, China

ICU = UTI

Para anorexia: p-value < 0.001
A probabilidade da diferença entre 67% e 30% ser APENAS pela aleatoriedade da amostragem é menor que 0.001. Ou seja, muito improvável.

https://towardsdatascience.com/how-to-interpret-p-value-with-covid-19-data-edc19e8483b

Inferência Estatística

- A inferência estatística não é mágica nem infalível, mas uma ferramenta extraordinária para explicar várias situações do mundo
- Podemos adquirir grande percepção de muitos fenômenos da vida apenas determinando a explicação mais provável

Você acha que esse cidadão bebeu exageradamente ou foi envenenado por um agente secreto russo?

Sugestão de estudo

- Capítulo 9 (Estatística, Charles Wheelan)
- Inferência - Portal Action