PROGRAMA DE
PÓS-GRADUAÇÃO EM
CIÊNCIA POLÍTICA | UFMG

DCP919B
Metodologia II

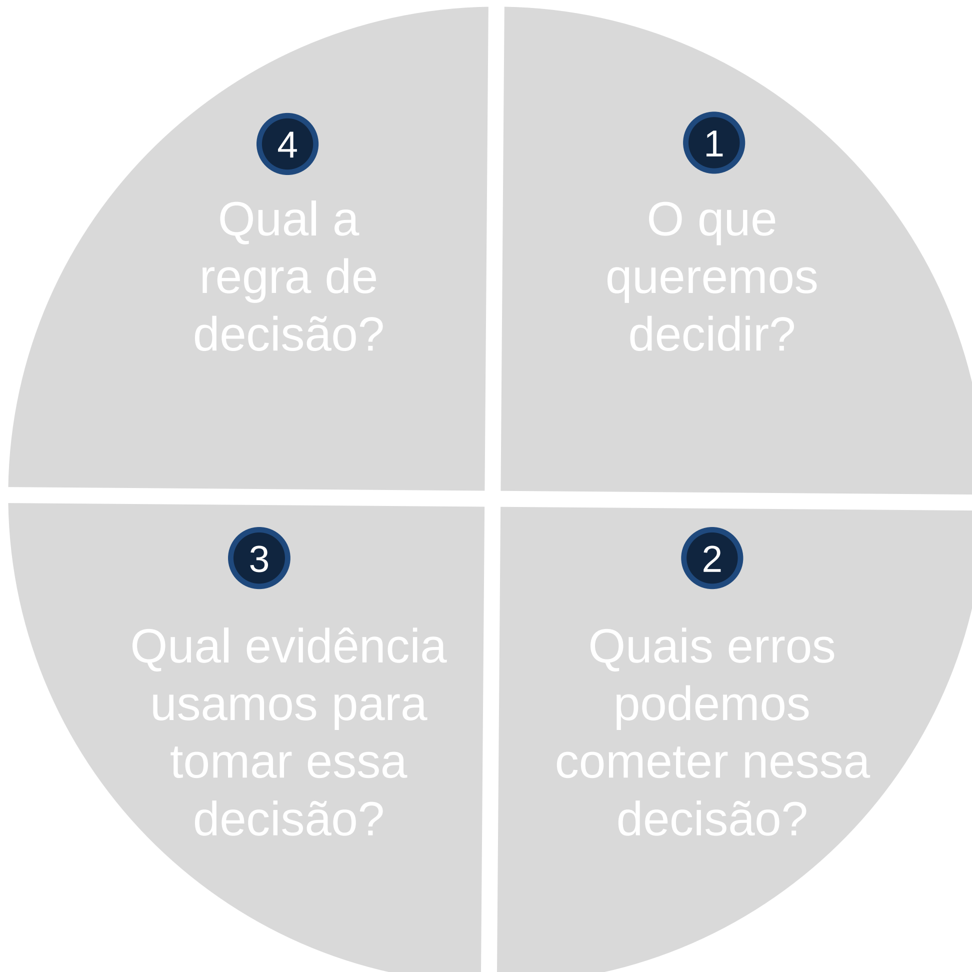
Teste de hipótese

Aula 4
28 de setembro de 2022

Ana Paula Karruz

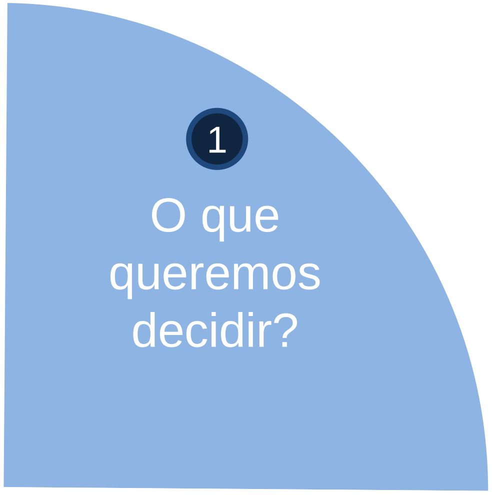
Conclusões sobre efeito de X em Y requerem tomada de decisão sobre o significado da evidência apurada

Teste de hipótese é rotina para essa tomada de decisão



Conclusões sobre efeito de X em Y requerem tomada de decisão sobre o significado da evidência apurada

Teste de hipótese é rotina para essa tomada de decisão



Como nos convencemos de que uma associação é real

(e não apenas fruto da aleatoriedade)

- **Testes de hipótese permitem avaliar se os dados observados são compatíveis com uma proposição** (e.g., donuts não afetam peso)

- Resultado de teste de hipótese **não é determinístico** (ou seja, resultado não é do tipo “donuts engordam” ou “donuts não engordam”)

- Ao contrário, resultado **expressa um grau de confiança** em relação à aderência entre a afirmação contida na hipótese e a evidência coletada (e.g., “com base nesta amostra, estamos 95% confiantes de que a afirmação ‘donuts não afetam peso’ é falsa”)

- No contexto de uma regressão, testes de hipótese informam, por exemplo, a **probabilidade de se ter obtido o $\beta_1$hat que obtivemos (dada sua imprecisão)**, em certo cenário

  – Cenário: previsto na hipótese testada (hipótese nula): H_0: $\beta_1 = 0$

  – Se a **probabilidade** de observar o **$\beta_1$hat (com respectivo grau de imprecisão)** que observamos for bem **baixa** dado o cenário de H_0, então **desconfiamos do cenário declarado em H_0**

Hipóteses testadas são **sobre parâmetros populacionais (βs)**, não sobre estimativas calculadas a partir da amostra (βhats)

A relação entre X e Y é "estatisticamente significante"?

Em outras palavras, a evidência suporta o cenário em que $\beta_j = 0$?

*As ferramentas estatísticas **não** nos permitem **provar [...] uma hipótese nula**. Em vez disso, **"rejeitamos"** ou **"deixamos de rejeitar"** as hipóteses nulas.*

Bailey (2016: 138)

*Quando **deixamos de rejeitar** uma hipótese nula, estamos dizendo que o **$\beta_1$hat** que observamos **não** seria particularmente **improvável** se a hipótese nula fosse verdadeira. Por exemplo, normalmente [não] rejeitamos a hipótese nula quando observamos um **pequeno $\beta_1$hat**. Esse resultado não seria nada surpreendente se $\beta_1 = 0$. Também podemos deixar de rejeitar hipóteses nulas quando a **incerteza é alta**. Ou seja, um **grande $\beta_1$hat pode não ser muito surpreendente** mesmo quando $\beta_1 = 0$ **se a variância de $\beta_1$hat for grande em relação ao valor de $\beta_1$hat**.*

Bailey (2016: 139)

*Se, no entanto, observarmos um grande $\beta_1$hat com um pequeno erro padrão, **rejeitaremos** a hipótese nula e nos referiremos ao **coeficiente como estatisticamente significante**.*

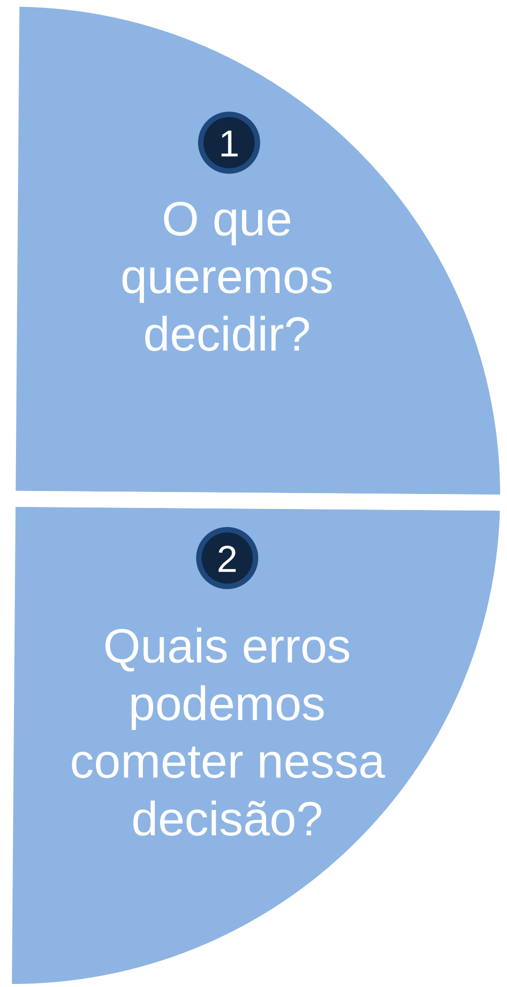
Bailey (2016: 138)

Quando rejeitamos uma hipótese nula, na verdade estamos dizendo que a probabilidade de se obter o $\beta_1$hat que estimamos é muito baixa se a hipótese nula fosse verdadeira.

Bailey (2016: 138)

Conclusões sobre efeito de X em Y requerem tomada de decisão sobre o significado da evidência apurada

Teste de hipótese é rotina para essa tomada de decisão



Nula (H_0) vs. Alternativa (H_1 ou H_a)

- **H_0: contém o cenário sendo testado**, e.g.:
  - H_0: $\beta_1 = 0$

- **H_1: contém todos os cenários concebíveis alternativos a H_0**, e.g.:
  - Hipótese alternativa bilateral (aka bicaudal):
    H_1: $\beta_1 \neq 0$
  - Hipótese alternativa direcional (aka unilateral ou unicaudal):
    H_1: $\beta_1 > 0$;
    outro exemplo: H_1: $\beta_1 < 0$

*Se **rejeitarmos a hipótese nula**, **aceitamos a hipótese alternativa**. **Não provamos** que a hipótese alternativa seja verdadeira. Em vez disso, a hipótese alternativa é a **ideia à qual nos apegamos** quando temos **evidência inconsistente** com a **hipótese nula**.*

Bailey (2016: 140)

O reconhecimento de que podemos incorrer em erro é central para o teste de hipótese

- Quando **rejeitamos H_0**, concluímos que o **cenário** nela contido (e.g., $\beta_1 = 0$) é **improvável dados o $\beta_1$hat** observado **e o se($\beta_1$hat)**. Nota: nossa conclusão não é que H_0 seja impossível

- Analogamente, quando **não rejeitamos H_0**, concluímos que o **cenário** nela contido (e.g., $\beta_1 = 0$) é **provável dados o $\beta_1$hat observado e o se($\beta_1$hat)**. Nota: nossa conclusão não é que H_0 seja verdadeira

- **Erros na conclusão de testes** de hipótese podem assumir 2 formas:

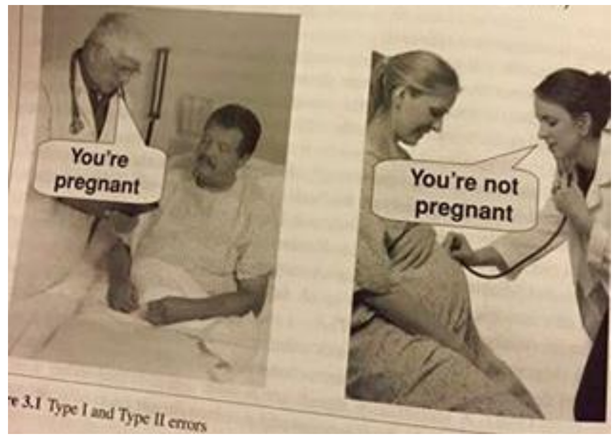
  – **Erro Tipo I:** Rejeitamos H_0 quando H_0 é verdadeira (e.g., saímos acreditando que X explica Y quando na verdade não explica); metáfora: julgar réu culpado quando ele não cometeu o crime

  – **Erro Tipo II:** Não rejeitamos H_0 quando H_0 é falsa (e.g., saímos acreditando que X não explica Y quando na verdade explica); metáfora: julgar réu inocente quando ele cometeu o crime

Erro Tipo II tende a ocorrer quando a amostra é pequena e/ou quando o nível de confiança (100-nível de significância) pré-definido para o teste é alto

Ilustração: erros Tipo I e Tipo II no teste de gravidez

H_0: gravidez = 0



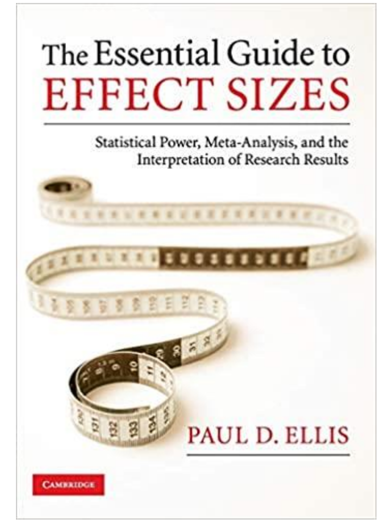


Tabela ou matriz de confusão (realmente, é assim que se chama!)

	Estado verdadeiro da natureza	
Decisão	**A hipótese nula é verdadeira**	**A hipótese nula é falsa**
Rejeitar a hipótese nula	**Erro tipo I** (rejeitar uma hipótese nula verdadeira); *α*	Decisão correta
Deixar de rejeitar a hipótese nula	Decisão correta	**Erro tipo II** (deixar de rejeitar uma hipótese nula falsa); *β*

Não é o mesmo β da equação de regressão

Atenção: Teste de hipótese não "prova" coisa alguma; incerteza não permitiria isso

(mesmo rejeitando H_0, incorremos em risco de erro Tipo I)

DECISÃO sobre hipótese nula	Implicação sobre hipótese alternativa
Rejeita	Aceita (mas não prova)
Não rejeita	**Não** rejeita

- Só mudamos nosso entendimento atual sobre a proposição testada (i.e., H_0) se pudermos falseá-la; teste não nos permite concluir que H_0 seja verdadeira
  - Encontrar evidência de que X afeta Y na amostra pode nos autorizar a concluir que X afeta Y na população, com um certo grau de confiança
  - Não encontrar evidência de que X afeta Y na amostra não nos autoriza a concluir que X não afeta Y na população (lembremos: a falta de evidência em favor de efeito pode advir da ausência de efeito na população ou da nossa incapacidade de detectá-lo)

Testes de hipótese focam no erro Tipo I, especificando a priori um nível aceitável (α) para esse erro

Quão improvável $\beta_j$hat (dada sua imprecisão) tem que ser para que rejeitemos H_0?

- O **nível de significância (α)** do teste, determinado a priori, especifica quão improvável $\beta_j$hat (com seu respectivo nível de imprecisão) tem que ser no cenário da H_0 para que rejeitemos essa hipótese

- Se α = 0,05, então rejeitamos H_0 se observarmos um $\beta_j$hat de magnitude tão grande (em relação ao erro padrão) que esperaríamos um $\beta_j$hat desse tamanho ou maior em no máximo 5% das amostras, se a nula for verdadeira

  – **α = probabilidade de rejeitar H_0 quando H_0 é verdadeira**
  – **α = risco assumido de cometer erro Tipo I**
  – **α = 100% – nível de confiança**

Atenção: à medida que diminuímos α, aumentamos a probabilidade de cometer erro Tipo II

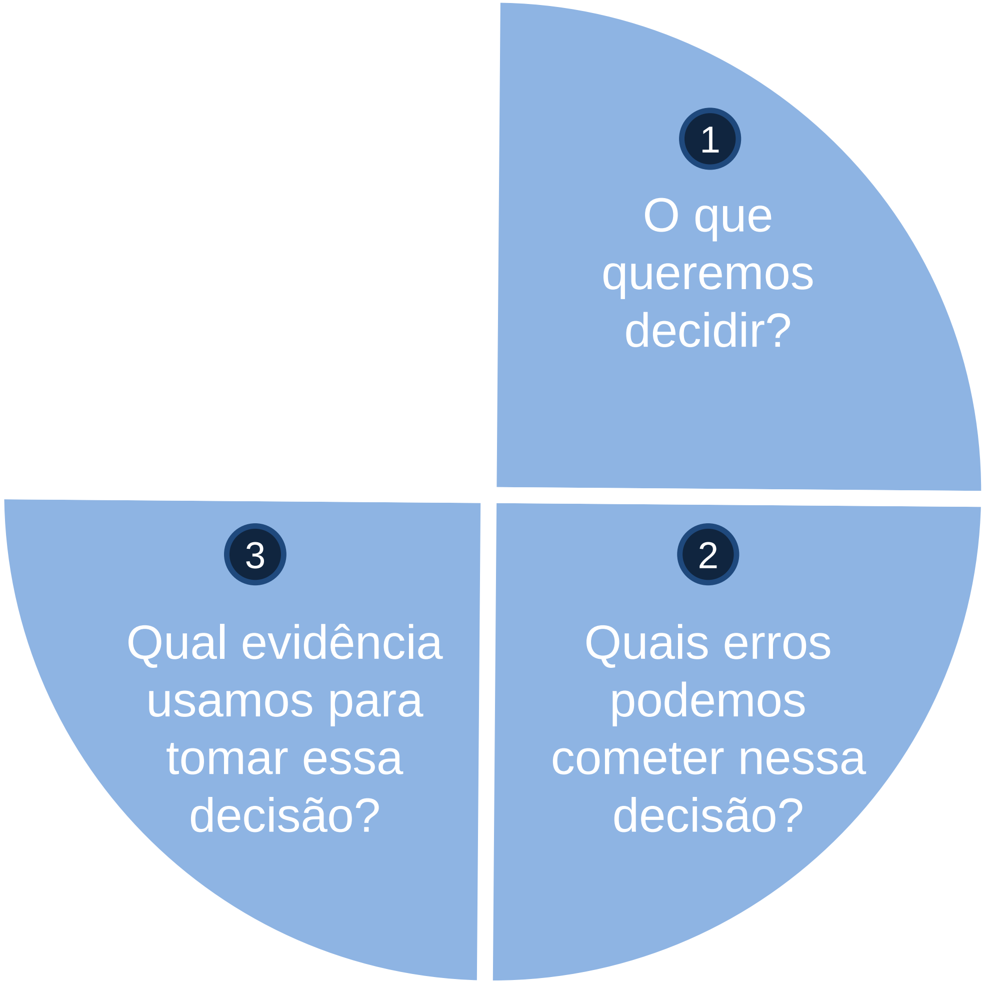
Níveis convencionais de significância (e confiança)

Nível de significância (α)	Nível de confiança
0,100 (10%)	90,0%
0,050 (5%)	95,0%
0,010 (1%)	99,0%
0,001 (0,1%)	99,9%

α = 0,10 (10%) é considerado nível de significância “marginal” (i.e., significância “café com leite”)

Conclusões sobre efeito de X em Y requerem tomada de decisão sobre o significado da evidência apurada

Teste de hipótese é rotina para essa tomada de decisão



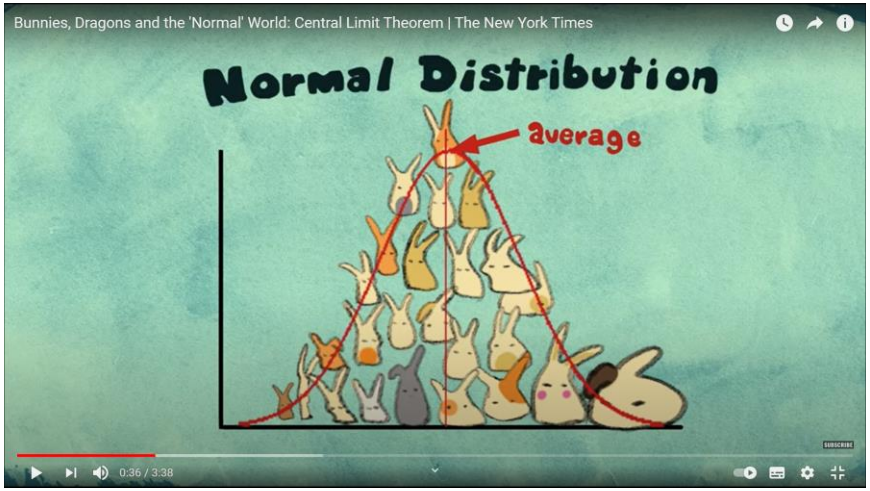
Distribuição de densidade de $\beta_j$hats se aproxima de uma normal

- **Fórmula** do $\beta_j$hat pode ser **reescrita como uma média** (Bailey, 2016: 85, nota 8)

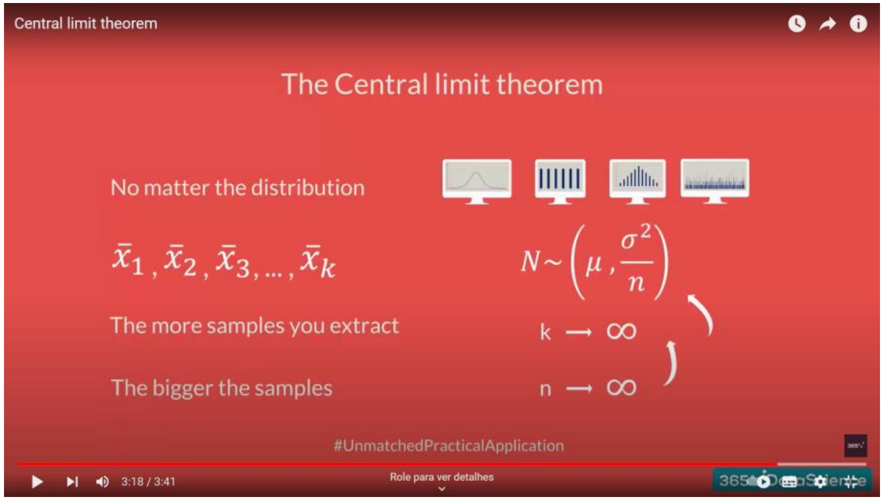
- Como tal, **$\beta_j$hat é uma estatística amostral sujeita ao Teorema Central do Limite** (aka Teorema do Limite Central), segundo o qual:
  - As médias amostrais de qualquer variável aleatória se distribuem de maneira próxima a uma distribuição normal centrada na média populacional (μ)

$$\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_k \sim N(\mu, \sigma^2/n)$$

  - Quanto maior o número de amostras (k), e quanto maior o tamanho de cada amostra (n), a distribuição amostral da média:
    - Será mais parecida com uma distribuição normal
    - Terá centro mais próximo da média populacional

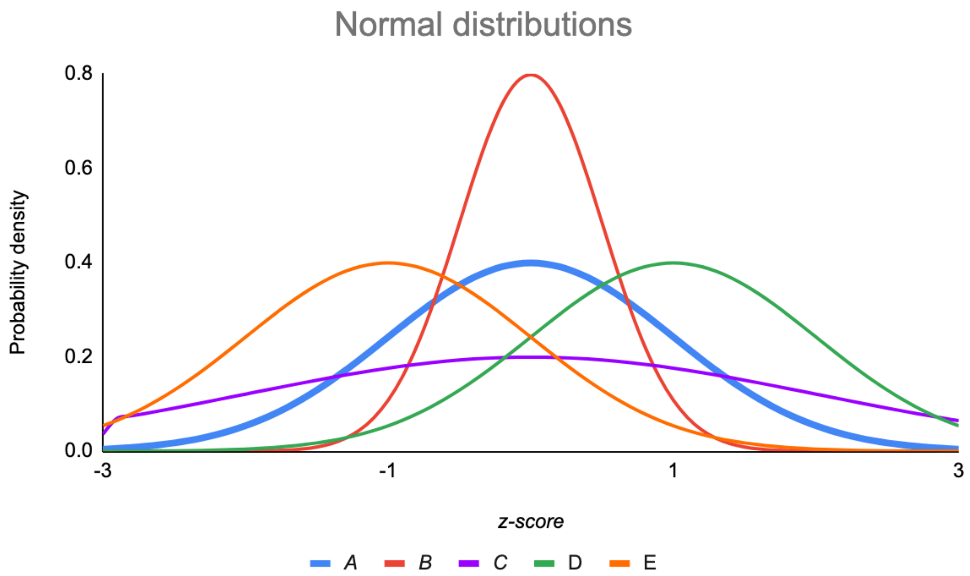


https://youtu.be/jvoxEYmQHNM



https://youtu.be/b5xQmk9veZ4
https://youtu.be/YAlJCEDH2uY

A distribuição $N(0, 1)$ é conhecida como "normal padrão" ou distribuição Z



Curve	Position or shape (relative to standard normal distribution)
A ($M = 0$, $SD = 1$)	Standard normal distribution
B ($M = 0$, $SD = 0.5$)	Squeezed, because $SD < 1$
C ($M = 0$, $SD = 2$)	Stretched, because $SD > 1$
D ($M = 1$, $SD = 1$)	Shifted right, because $M > 0$
E ($M = -1$, $SD = 1$)	Shifted left, because $M < 0$

Fonte: https://www.scribbr.com/statistics/standard-normal-distribution/

- Para **"padronizar" uma distribuição normal** (i.e., fazê-la ter média nula e desvio padrão unitário), aplicamos a seguinte fórmula:

$$z = \frac{x - \mu}{\sigma}$$

em que:
x = qualquer valor da distribuição normal original (aquela que se quer padronizar)
μ = média da distribuição normal original
σ = desvio padrão da distribuição normal original

- Um **z-score** (valor de Z para um dado valor de X) positivo significa que x é maior que a média de X
  - z-score > 0 ↔ x > μ
  - z-score < 0 ↔ x < μ
  - z-score = 0 ↔ x = μ

- Entre outras vantagens, **converter** uma distribuição normal **em uma distribuição normal padrão permite**:
  - Comparar valores de diferentes distribuições com diferentes médias e desvios padrão
  - Encontrar rapidamente (numa tabela) a probabilidade de observações em uma distribuição caírem acima ou abaixo de um determinado valor

- Quando os parâmetros populacionais μ e σ são desconhecidos, é possível estimar o z-score a partir das estatísticas amostrais $\bar{x}$ (média amostral de X) e s (desvio padrão amostral de X). **Alerta:** s tende a subestimar σ

$$z_{estimado} = \frac{x - \bar{x}}{s}$$

Estamos tentando descobrir se $\beta_j$hat seria altamente improvável se H_0 for verdadeira

Desafio: Como $\beta_j$hat tem a escala do Y, a escala de $\beta_j$hat poderia ser qualquer uma!

*O que nos importa não é o coeficiente estimado $\beta_1$hat em si, mas **quão grande** o coeficiente **$\beta_1$hat é em relação a seu erro padrão**.*

Bailey (2016: 148)

*Portanto usamos uma **estatística de teste** que consiste no coeficiente estimado dividido pelo seu erro padrão estimado:*

$$\frac{\hat{\beta}_j}{EP(\hat{\beta}_j)}$$

Esta razão é chamada de **estatística t**; tem **mesmo sinal de $\beta_j$hat**, pois se($\beta_j$hat) > 0

Estamos padronizando a distribuição normal de β1hat, porém usando o erro padrão (calculado a partir da amostra) no lugar do desvio padrão populacional

Podemos testar qualquer hipótese do tipo **H_0: $\beta_j = \beta_j^{Nula}$** usando a seguinte razão como estatística de teste:

$$\frac{\hat{\beta}_j - \beta_j^{Nula}}{EP(\hat{\beta}_j)}$$

*Assim, **nossa estatística de teste**, que corresponde a **nossa evidência sobre β_j**, reflete **quantos erros padrão acima ou abaixo de zero [ou outro valor contido em H_0] o coeficiente estimado está**.*

Bailey (2016: 148)

Enquanto H_0 define uma média para a distribuição de $\beta_j$hats, recorremos à amostra para estimar o erro padrão

Qual é a distribuição de $\beta_1$hat ***sob a hipótese nula*** [H_0: $\beta_1 = 0$]*? Muito simples: É uma variável aleatória normalmente distribuída e* ***centrada*** *em zero porque MQO é um estimador não enviesado* [dado que suas premissas tenham sido atendidas] *e, se o verdadeiro valor de β_1 é zero* [sob a hipótese nula]*, então uma distribuição não enviesada de $\beta_1$hat estará centrada em zero.*

Bailey (2016: 141)

Quão larga [espalhada] *é a distribuição de $\beta_1$hat sob a hipótese nula? Em contraste com a média da distribuição [média esta] que assumimos sob a nula, a* ***largura depende dos dados e do erro padrão*** *calculado com base nesses dados. Em outras palavras, permitimos aos dados que nos informem sobre o erro padrão de $\beta_1$hat* […]*.*

Bailey (2016: 142)

A distribuição t

aka t de Student

Quando os tamanhos das amostras são pequenos, a distribuição t é usada como alternativa à distribuição normal para estimar a confiança ou determinar valores críticos [...]. Essa substituição ocorre porque o desvio padrão amostral [assim como o erro padrão] é uma medida enviesada do desvio padrão populacional, e tende a subestimá-lo.

Brereton (2015: 481)
https://doi.org/10.1002/cem.2713

A distribuição t

aka t de Student

t(ν)

ν = letra grega
Nu minúscula

- A função de densidade de probabilidade é **simétrica** e sua forma se assemelha à forma de **sino** de uma distribuição **normal padrão**, com a diferença que t **possui caudas mais pesadas**
- A média e o **centro** da distribuição são **0**
- A **variância é sempre maior que 1**, embora seja próxima de 1 quando há muitos graus de liberdade
- **Com infinitos graus de liberdade, a distribuição t é a mesma que a distribuição normal padrão (Z)**. Por esta razão, ν também é conhecido como parâmetro de normalidade

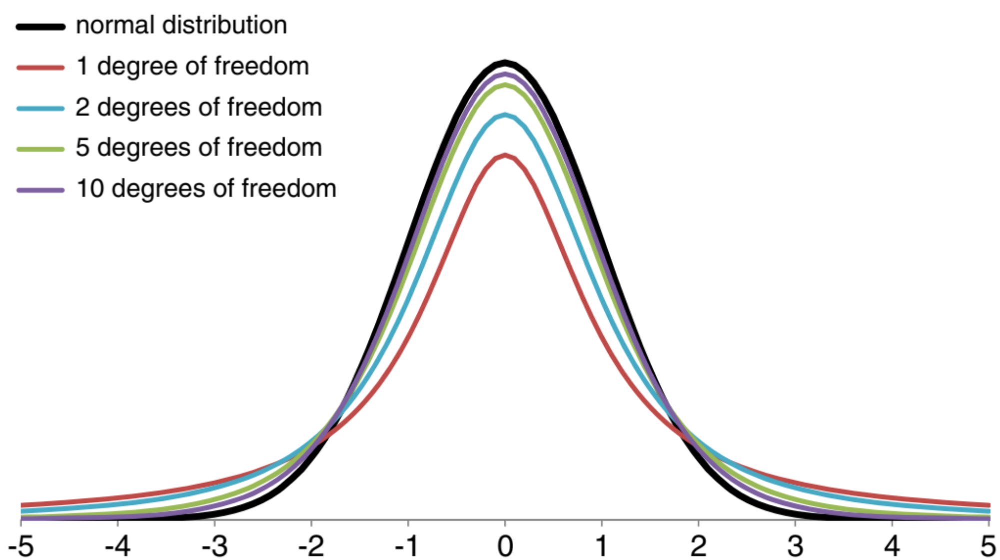


Figure 1. The probability density function for the normal distribution together with the *t*-distributions for different degrees of freedom. The horizontal axis represents standard deviations: for the normal distribution, this is the population standard deviation and for the *t*-distribution, the sample standard deviation.

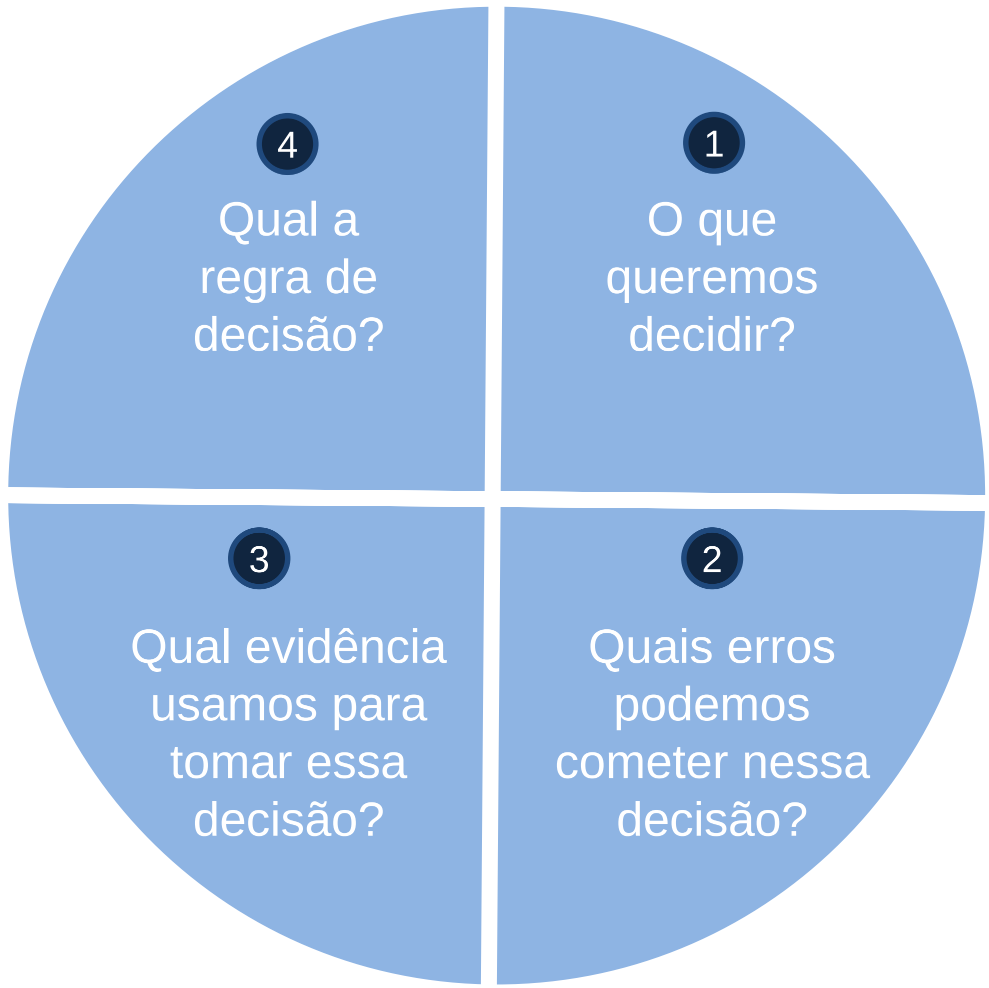
Brereton (2015: 482)
https://doi.org/10.1002/cem.2713

Apêndice
DEEP DIVE 1:
Por que estatística t apresenta distribuição t_{n-k-1}?

Não será cobrado

Conclusões sobre efeito de X em Y requerem tomada de decisão sobre o significado da evidência apurada

Teste de hipótese é rotina para essa tomada de decisão



À medida que o tamanho da amostra cresce, a distribuição t tende à normal padrão

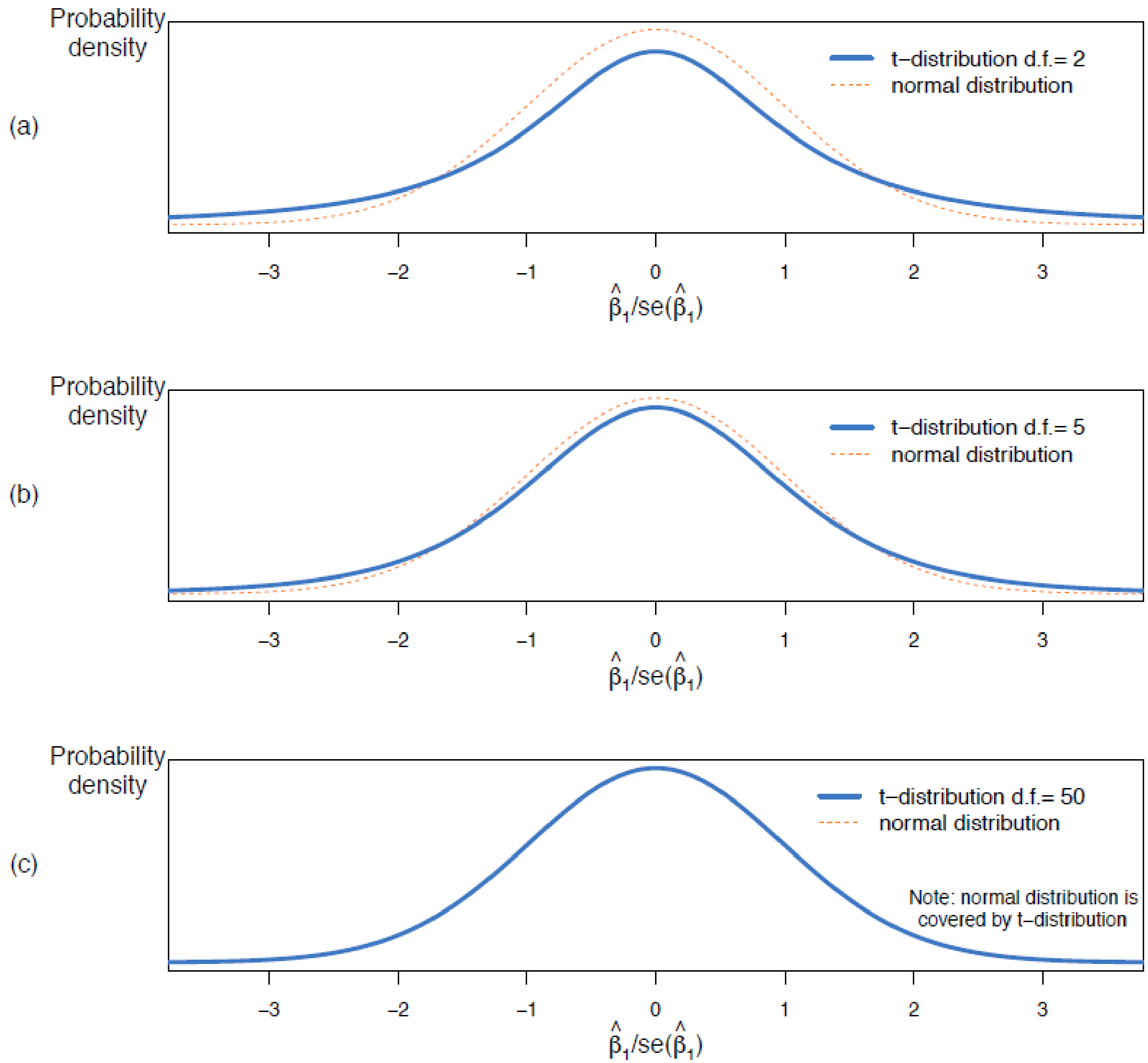
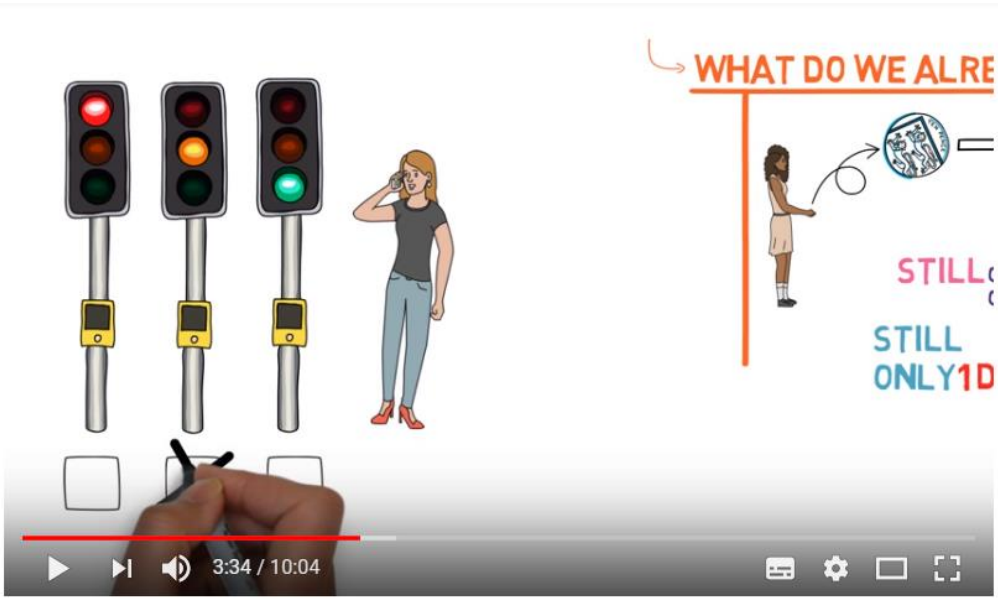


FIGURE 4.3: Three t distributions

Fonte: Bailey (2016: 152).

Como ilustrado na página anterior, a forma específica da distribuição t depende dos graus de liberdade (n-k-1)



what are degrees of freedom?

https://youtu.be/rATNoxKg1yA

Graus de liberdade: número de observações independentes utilizadas para calcular uma estatística

Intuição: Graus de liberdade

Dados o conjunto *D* e sua média ($\overline{D}$):
$D = \{2, 4, 6, 8, 10\}$
$\overline{D} = 6$

Se conheço $\overline{D}$ e *N*-1 elementos de *D*, então posso determinar o n-ésimo elemento de *D*, sem qualquer dúvida sobre seu valor.

$$\overline{D} = \frac{2+4+6+8+d}{(N-1)+1} = 6$$

$$30 = 20 + d$$
$$d = 30 - 20 = 10$$

Diferente dos elementos até *N*-1, o n-ésimo elemento não é “livre”, pois só poderia admitir um certo valor, dado o que já sei sobre *D* (ou seja, dado que conheço $\overline{D}$).

֍֍֍

Note que no cálculo de **$\beta_1$hat** (coeficiente de inclinação estimado em uma regressão simples) precisamos de $\bar{X}$ e $\bar{Y}$ para obter os desvios de cada observação em relação à média dessas variáveis. Assim, a média dessas variáveis é conhecida antes de calcularmos o $\hat{\beta}_1$. Portanto, a observação (X_N, Y_N) não é “livre”.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

O valor crítico é o limiar a partir do qual a estatística t [ou seja, $\beta_j$hat / se($\beta_j$hat)] é tão improvável que rejeitamos H_0

- O **valor crítico depende** de:
  - Graus de liberdade: **n – k – 1; k = número de variáveis explicativas**
  - Formato da hipótese alternativa: **H_1 unilateral ou bilateral**
  - Nível de significância estatística: **α**

Qual é a H_0 em cada um desses painéis?

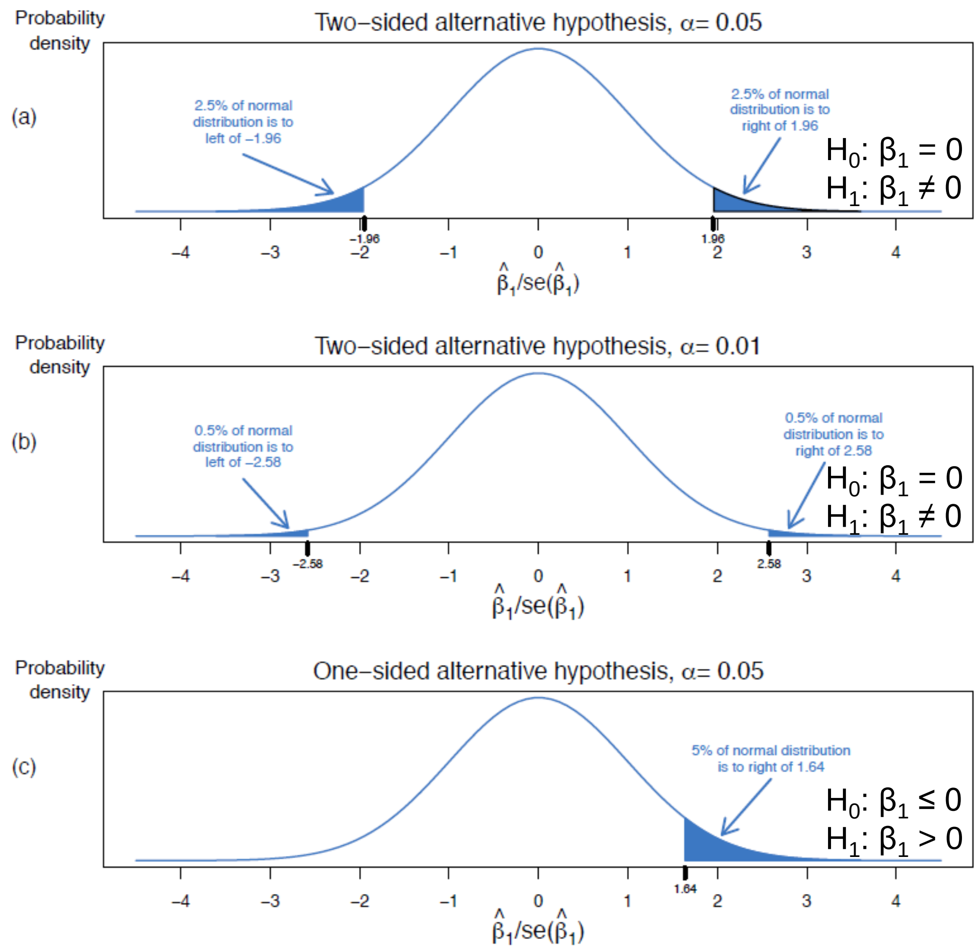
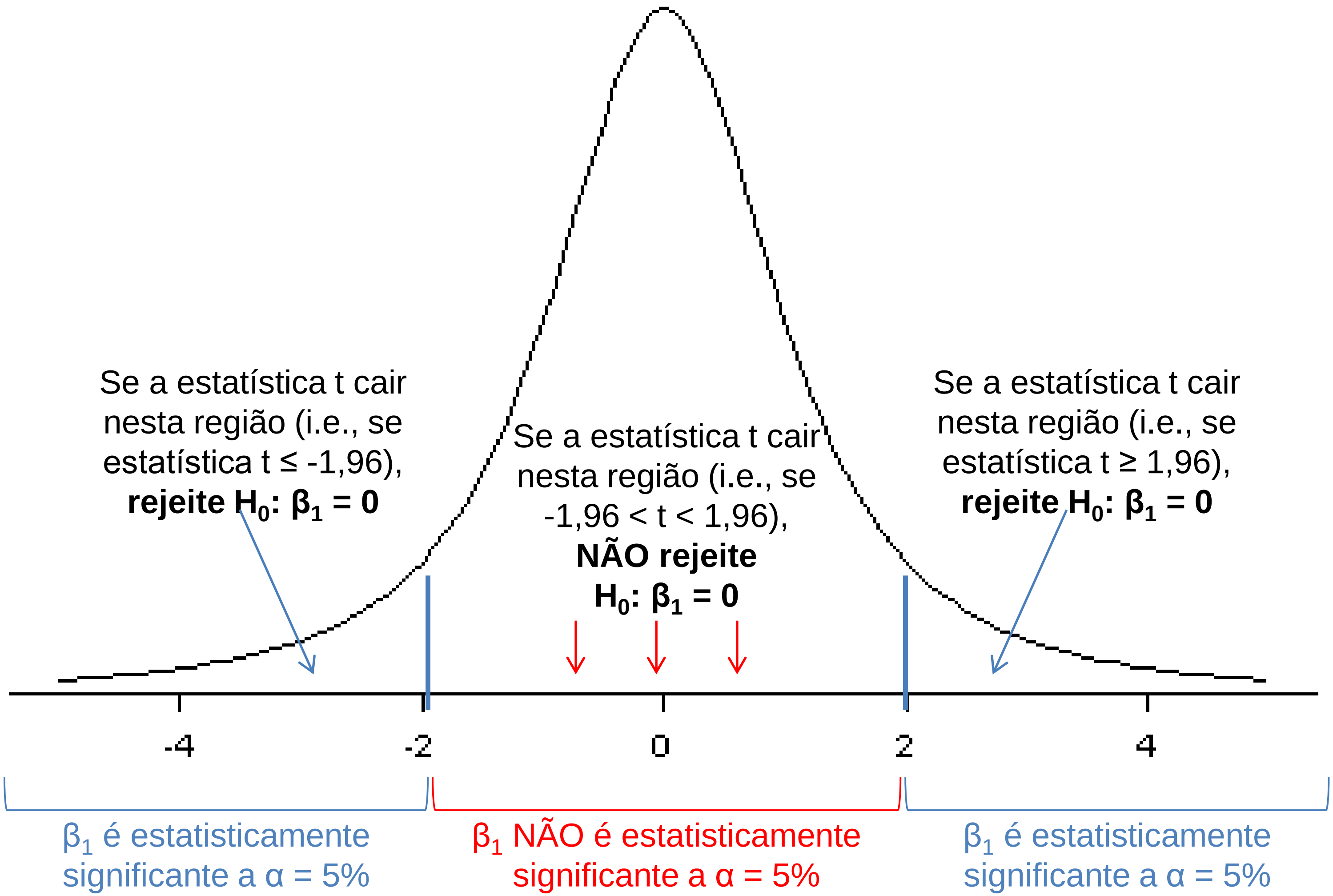


FIGURE 4.4: Critical Values for Large Sample t tests

Fonte: Bailey (2016: 155).

Zooming: áreas de rejeição para H_0: $\beta_1 = 0$ (teste bilateral com amostra grande e α = 0,05)



One- or two-sided?

Better go for two!

Observe que o valor crítico unilateral para α = 0,05 é menor que o valor crítico bilateral. ***Os valores críticos unilaterais serão sempre menores para qualquer valor de α****, o que significa que é* ***mais fácil rejeitar a hipótese nula para uma hipótese alternativa unilateral*** *do que para uma hipótese alternativa bilateral. Portanto, usar valores críticos com base em uma* ***alternativa bilateral*** *é* ***estatisticamente cauteloso*** *no sentido de que é menos provável que pareçamos ansiosos demais para rejeitar a nula se usarmos uma alternativa bilateral.*

Bailey (2016: 156)

Exceto se indicado em contrário, nesta disciplina trataremos de testes t bicaudais

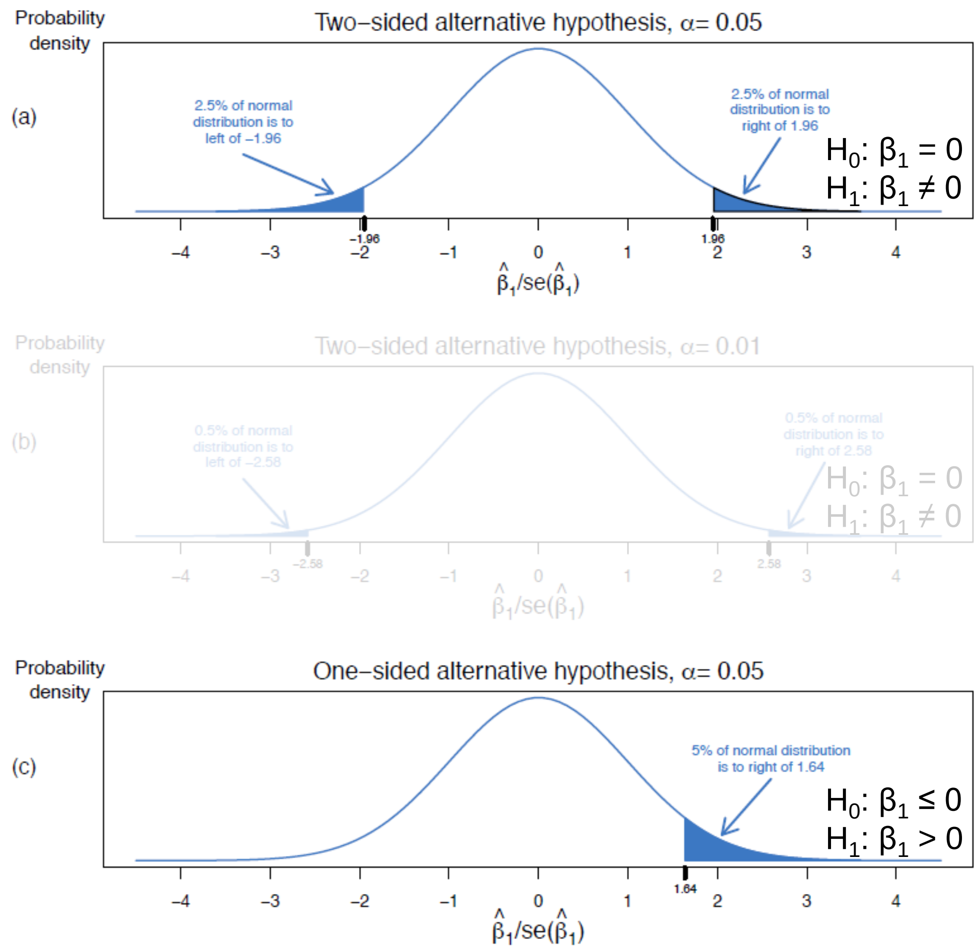


FIGURE 4.4: Critical Values for Large Sample t tests

Fonte: Bailey (2016: 155).

Valores críticos diminuem à medida que graus de liberdade aumentam

- Quanto **menor** o tamanho da **amostra**, **mais incerteza** temos sobre o **$\beta_j$hat** e portanto **mais elevado** será o **valor crítico**

[dito de outra forma...]

- Quanto **maior** o tamanho da **amostra**, **menos incerteza** temos sobre o **$\beta_j$hat** e portanto **menos elevado** será o **valor crítico**

Table 4.4: Critical Values for t distribution

	α (1-sided) $\Rightarrow$	0.05	0.025	0.01	0.005
	α (2-sided) $\Rightarrow$	0.10	0.050	0.02	0.01
Degrees of freedom	2	2.92	4.30	6.97	9.92
	5	2.01	2.57	3.37	4.03
	10	1.81	2.23	2.76	3.17
	15	1.75	2.13	2.60	2.95
	20	1.73	2.09	2.53	2.85
	50	1.68	2.01	2.40	2.68
	100	1.66	1.98	2.37	2.63
	∞	1.64	1.96	2.32	2.58

A t distribution with ∞ degrees of freedom is the same as a normal distribution.

Fonte: Bailey (2016: 157).

*À medida que os graus de liberdade aumentam, a distribuição t se parece **cada vez mais com uma distribuição normal [padrão]** e, para infinitos graus de liberdade, é exatamente como uma distribuição normal [padrão], produzindo valores críticos idênticos. Para **graus de liberdade acima de 100**, é razoável usar valores críticos da **distribuição normal [padrão]** como uma boa aproximação.*

Bailey (2016: 156-157)

Resumão (até agora)

We compare $\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$ to our critical value and reject H_0 if the magnitude is larger than the critical value. We refer to the ratio of $\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$ as the **t statistic** (or, "t stat" as the kids say). The t statistic is so named because that ratio will be compared to a critical value that depends on the t distribution in the manner we have just outlined. Tests based on two-sided alternative tests with $\alpha = 0.05$ are very common. When the sample size is large, the critical value for such a test is 1.96. Hence the rule of thumb is that a t statistic bigger than 2 **indicates that β_1** is statistically significant at conventional levels.

Fonte: Bailey (2016: 157).

Temos evidência suficiente para acreditar que donuts engordam?

Vide aula 02

Interpretação dos coeficientes estimados

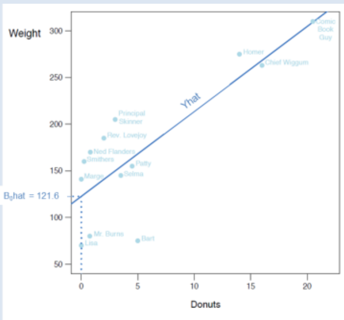


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield

Fonte: Adaptado de Bailey (2016, p 9).

1 libra = 0,454 quilograma
1 quilograma = 2,205 libras

- Estimação da linha de regressão fornece os seguintes resultados:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 * X_i$$

$$\hat{Y}_i = 121.6 + 9.2 * X_i$$

  onde Y = peso em libras, X = donuts por semana, e i é o subscrito que indexa indivíduos

- **Intercepto (constante):** Estima-se que indivíduos que não consomem donuts pesem 121,6 pounds, em média

- **Coeficiente de inclinação:** Estima-se que o consumo de um donut adicional por semana esteja associado a um aumento 9,2 pounds no peso

ceteris paribus?

```
> summary(reg.pounds)

Call:
lm(formula = dados$Weight..pounds. ~ dados$Donuts.per.week)

Residuals:
    Min      1Q  Median      3Q     Max 
-92.731 -13.508   3.916  36.081  55.716 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            121.613     16.593   7.329 1.49e-05 ***
dados$Donuts.per.week    9.224      1.959   4.707 0.000643 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.81 on 11 degrees of freedom
Multiple R-squared:  0.6683,	Adjusted R-squared:  0.6381 
F-statistic: 22.16 on 1 and 11 DF,  p-value: 0.0006426

> qt(1-0.05/2, 11)
[1] 2.200985
```

Para um teste unilateral, utilizaríamos o seguinte comando para obter o t crítico: `qt(1-0.05, 11)`

Passo a passo do teste t de significância estatística de β_1

1) Especifique as **hipóteses**
   H_0: $\beta_1 = 0$
   H_1: $\beta_1 \neq 0$

2) Apure o valor crítico (ou $\mathbf{t_{c(GdeL,\alpha)}}$, o chamado **t crítico**), considerando n – k – 1 = 11 e α = 0,05
   Vide tabela t, ou aplique o seguinte comando em R:
   `qt(1-0.05/2, 11)`
   $t_c = 2{,}201$

3) Calcule a **estatística t**
   t = $\beta_1$hat / se($\beta_1$hat)
   t = 9,224 / 1,959 = 4,709

4) Aplique a **regra de decisão para teste bilateral**
   Se $|t| \geq t_c$, então rejeite H_0
   Se $|t| < t_c$, então não rejeite H_0
   |4,709| > 2,201 ➔ Rejeite H_0

Rejeitamos H_0: $\beta_1 = 0$ em favor de H_1, ao nível de significância de 5%. Encontramos na amostra evidência suficiente para acreditar que $\beta_1 \neq 0$ e que, portanto, donuts afetam o peso. Essa associação é considerada estatisticamente significante (i.e., estatisticamente diferente de zero), com 95% de confiança. Dada a significância estatística de β_1 e dado que $\beta_1$hat > 0, consumo de donuts parece ter um efeito positivo sobre peso.

OK, rejeitamos H_0 a 5% de significância. Qual o menor nível de significância ao qual rejeitaríamos H_0?

Vide aula 02

Interpretação dos coeficientes estimados

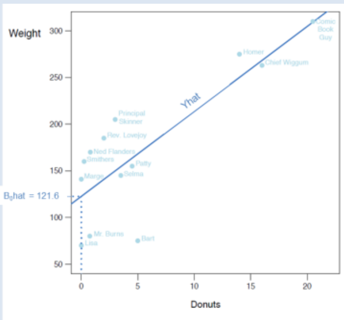


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield

Fonte: Adaptado de Bailey (2016, p 9).

- Estimação da linha de regressão fornece os seguintes resultados:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 * X_i$$

$$\hat{Y}_i = 121.6 + 9.2 * X_i$$

onde Y = peso em libras, X = donuts por semana, e i é o subscrito que indexa indivíduos

- **Intercepto (constante):** Estima-se que indivíduos que não consomem donuts pesem 121,6 pounds, em média
- **Coeficiente de inclinação:** Estima-se que o consumo de um donut adicional por semana esteja associado a um aumento 9,2 pounds no peso

ceteris paribus?

1 libra = 0,454 quilograma
1 quilograma = 2,205 libras

```
> summary(reg.pounds)

Call:
lm(formula = dados$Weight..pounds. ~ dados$Donuts.per.week)

Residuals:
    Min      1Q  Median      3Q     Max 
-92.731 -13.508   3.916  36.081  55.716 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            121.613     16.593   7.329 1.49e-05 ***
dados$Donuts.per.week    9.224      1.959   4.707 0.000643 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.81 on 11 degrees of freedom
Multiple R-squared:  0.6683,	Adjusted R-squared:  0.6381 
F-statistic: 22.16 on 1 and 11 DF,  p-value: 0.0006426

> 2*pt(-abs(4.709),df=11)
[1] 0.0006407414
> pt(abs(4.709),df=11)
[1] 0.9996796
```

Num teste bicaudal, p-valor é a probabilidade de se obter uma estatística t de magnitude ≥ à estatística t observada se H_0 for verdadeira

- Para β_1, p-valor = Pr(>|t|) = 2*Pr(t>|4,707|) = 0,000643
- Esse é um p-valor bastante baixo, implicando que $\beta_1$hat (com o grau de imprecisão a ele associado) é realmente muito improvável se H_0 for verdadeira
  - p-valor pequeno fornece evidência contra H_0
  - p-valor alto fornece pouca evidência contra H_0
- Podemos conduzir um teste de hipótese com base no p-valor, aplicando a seguinte regra de decisão:
  Se p-valor ≤ α, então rejeite H_0
  Se p-valor > α, então não rejeite H_0

  0,000643 < 0,05 ➔ Rejeite H_0

Além de indicar se rejeitaremos H_0, o **p-valor informa o nível mínimo de significância ao qual poderíamos rejeitar H_0**; assim, o p-valor informa o peso da evidência a favor de H_0. Neste exemplo, β_1 é estatisticamente significante (i.e., estatisticamente diferente de zero) a 0,1% (de fato, é significante a 0,0643%). Portanto, β_1 é também estatisticamente significante aos níveis convencionais de significância: 1%, 5% e 10% (significância marginal).

Significant other

AnnaSophie Robb e Dwayne Johnson em “Race to Witch Mountain” (“A Montanha Enfeitiçada”), 2009.

Amor ou cilada?

A Statistically Significant Love Song

https://youtu.be/tVx2V75hWRY

Uma forma de reconhecer a aleatoriedade de $\beta_j$hat é oferecer uma estimativa intervalar de β_j

- **Intuição:** podemos representar nosso grau de confiança sobre o valor do verdadeiro β_j apresentando um intervalo que, acreditamos, contenha esse coeficiente populacional

- **O intervalo de confiança define uma faixa de valores de β_j que seriam os mais consistentes com a evidência (i.e., com $\beta_j$hat e seu respectivo grau de imprecisão)**

  – Para os valores de β_j contidos no intervalo de confiança, **nosso $\beta_j$hat e respectivo se($\beta_j$hat) não seriam surpreendentes** (i.e., não seriam improváveis)

  – Escolhemos o que entendemos por “improvável” ao definir um nível de significância (α) ou, alternativamente, um nível de confiança (100% – α)

- O intervalo de confiança oferece uma **estimativa intervalar** de β_j, enquanto **$\beta_j$hat** é nossa **estimativa pontual** de β_j

Uma forma de reconhecer a aleatoriedade de $\beta_1$hat é oferecer uma estimativa intervalar de β_1

R

Vide aula 02

Interpretação dos coeficientes estimados

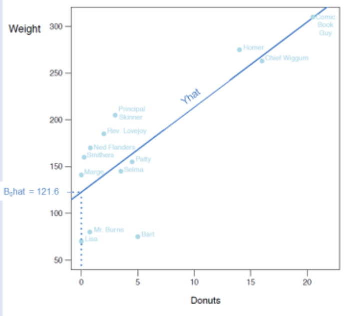


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield

Fonte: Adaptado de Bailey (2016, p 9).

- Estimação da linha de regressão fornece os seguintes resultados:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 * X_i$$

$$\hat{Y}_i = 121.6 + 9.2 * X_i$$

onde Y = peso em libras, X = donuts por semana, e i é o subscrito que indexa indivíduos

- **Intercepto (constante):** Estima-se que indivíduos que não consomem donuts pesem 121,6 pounds, em média
- **Coeficiente de inclinação:** Estima-se que o consumo de um donut adicional por semana esteja associado a um aumento 9,2 pounds no peso

ceteris paribus?

1 libra = 0,454 quilograma
1 quilograma = 2,205 libras

```
> summary(reg.pounds)
[...]
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            121.613     16.593   7.329 1.49e-05 ***
dados$Donuts.per.week    9.224      1.959   4.707 0.000643 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[...]

> confint(reg.pounds)
                          2.5 %    97.5 %
(Intercept)           85.092114 158.13471
dados$Donuts.per.week  4.910836  13.53622

> qt(1-0.01/2, 11)
[1] 3.105807

> confint(reg.pounds, level = .99)
                          0.5 %    99.5 %
(Intercept)           70.078268 173.14856
dados$Donuts.per.week  3.137897  15.30916
```

- O intervalo de confiança está **centrado em $\beta_1$hat**
- Sua **extensão** dependente do erro padrão de $\beta_1$hat, dos graus de liberdade e do nível de confiança

IC = $\beta_1$hat ± [se($\beta_1$hat) * t_c]

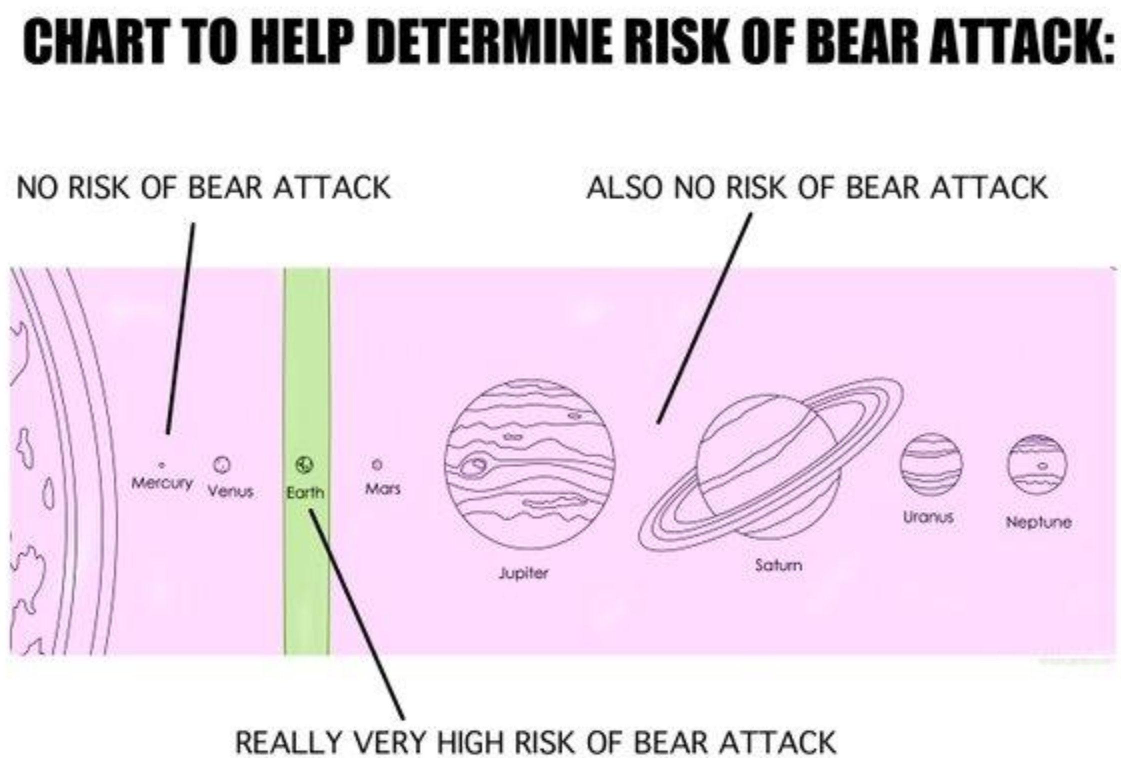
- **Para α = 0,05 (i.e., nível de confiança = 95%):**
  Limite inferior = 9,224 – (1,959 * 2,201) = 4,912
  Limite superior = 9,224 + (1,959 * 2,201) = 13,536
  $4,912 \leq \beta_1 \leq 13,536$

  Interpretação rigorosa: Se repetíssemos o estudo muitas e muitas vezes, com amostras de mesmo tamanho, aproximadamente 95% das estimativas intervalares conteriam o verdadeiro parâmetro populacional

  Interpretação coloquial: Estamos 95% confiantes que o verdadeiro β_1 esteja contido no intervalo de confiança

- **Para α = 0,01 (i.e., nível de confiança = 99%):**
  Limite inferior = 9,224 – (1,959 * 3,106) = 3,139
  Limite superior = 9,224 + (1,959 * 3,106) = 15,309
  $3,139 \leq \beta_1 \leq 15,309$

Quanto maior o nível de confiança, mais largo o intervalo de confiança.
Quais os limites de um IC com 100% de confiança?



(Very accurate, but not precise.)

Fonte: Desconhecida.

Uma forma de reconhecer a aleatoriedade de $\beta_j$hat é oferecer uma estimativa intervalar de β_j

Como o **intervalo de confiança** informa a gama de valores de β_j consistentes com a evidência, **também nos indica se rejeitaríamos H_0.**

Se esse **intervalo não incluir zero**, então zero não seria um valor de β_j que provavelmente produziria os dados e as estimativas que observamos e, portanto, **rejeitaríamos H_0: $\beta_j = 0$**, ao nível de significância α.

Apêndice
DEEP DIVE 2:
Qual a intuição por trás do intervalo de confiança?

Não será
cobrado

Testes de hipótese apoiam a inferência estatística (extração de conclusões a partir da amostra), mas têm limitações

Testes de significância estatística...

- Nada têm a ver com a **solidez teórica** do modelo

- São baseados na premissa de que **não há endogeneidade**

- Podem conduzir a **diferentes conclusões para cenários virtualmente iguais**: |t| apenas ligeiramente superior (inferior) ao t crítico

- Podem sugerir **conclusões idênticas para cenários dramaticamente diferentes**: |t| muito (apenas ligeiramente) superior ao t crítico

p-valor ajuda aqui

- São afetados pelo tamanho da amostra: **quanto maior a amostra, maior a chance de encontrar significância estatística**. Isto ocorre porque o t crítico é menor para amostras maiores; também, o erro padrão tende a ser menor quanto maior for o tamanho da amostra
  - Analogamente, em **amostras pequenas**, o erro padrão tende a ser alto, de forma que o teste pode **não rejeitar H_0**, ainda que $\beta_j$hat tenha magnitude expressiva

- Não indicam **significância material**: β_j pode ser estatisticamente significante ainda que sua magnitude seja desprezível em termos práticos

Para saber mais sobre p-valor e limites dos testes de hipótese

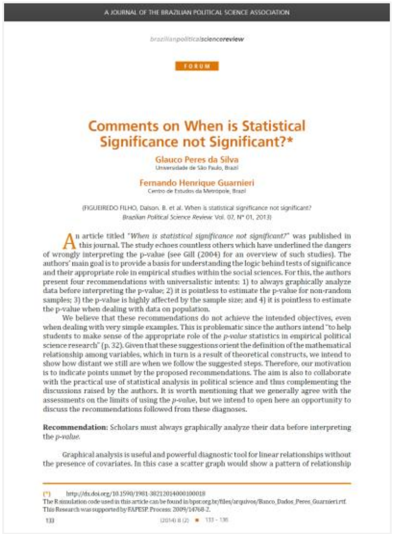

brazilianpoliticalsciencereview

ARTICLE

When is statistical significance not significant?

Dalson Britto Figueiredo Filho
Political Science Department, Federal University of Pernambuco (UFPE), Brazil
Ranulfo Paranhos
Social Science Institute, Federal University of Alagoas (UFAL), Brazil
Enivaldo C. da Rocha
Political Science Department, Federal University of Pernambuco (UFPE), Brazil
Mariana Batista
Ph.D candidate in Political Science, Federal University of Pernambuco (UFPE), Brazil
José Alexandre da Silva Jr.
Social Science School, Federal University of Goiás (UFG), Brazil
Manoel L. Wanderley D. Santos
Department of Political Science, Federal University of Minas Gerais (UFMG), Brazil
Jacira Guiro Marino
Carlos Drummond de Andrade School (FCDA), Brazil

Figueiredo Filho et al. (2013)


A JOURNAL OF THE BRAZILIAN POLITICAL SCIENCE ASSOCIATION

brazilianpoliticalsciencereview

FORUM

Comments on When is Statistical Significance not Significant?*

Glauco Peres da Silva
Fernando Henrique Guarnieri

Silva e Guarnieri (2014)


A JOURNAL OF THE BRAZILIAN POLITICAL SCIENCE ASSOCIATION

brazilianpoliticalsciencereview

FORUM

Reply on the Comments on When is Statistical Significance not Significant?

Dalson B. Figueiredo Filho
Universidade Federal de Pernambuco, Brazil
Enivaldo C. da Rocha
Universidade Federal de Pernambuco, Brazil
Mariana Batista
Universidade Federal de Pernambuco, Brazil
Ranulfo Paranhos
Universidade Federal de Alagoas, Brazil
José Alexandre da Silva Jr.
Universidade Federal de Alagoas, Brazil

Figueiredo Filho et al. (2014)


THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

Q: Why do so many colleges and grad schools teach $p = 0.05$?
A: Because that's still what the scientific community and journal editors use.
Q: Why do so many people still use $p = 0.05$?
A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do; we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

The ASA Board was also stimulated by highly visible discussions over the last few years. For example, ScienceNews (Siegfried 2010) wrote: "It's science's dirtiest secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation." A November 2013, article in Phys.org Science News Wire (2013) cited "numerous deep flaws" in null hypothesis significance testing. A ScienceNews article (Siegfried 2014) on February 7, 2014, said "statistical techniques for testing hypotheses …have more flaws than Facebook's privacy policies." A week later, statistician and "Simply Statistics" blogger Jeff Leek responded. "The problem is not that people use P-values poorly," Leek wrote, "it is that the vast majority of data analysis is not performed by people properly trained to perform data analysis" (Leek 2014). That same week, statistician and science writer Regina Nuzzo published an article in *Nature* entitled "Scientific Method: Statistical Errors" (Nuzzo 2014). That article is now one of the most highly viewed *Nature* articles, as reported by altmetric.com (*http://www.altmetric.com/details/2115792#score*).

Of course, it was not simply a matter of responding to some articles in print. The statistical community has been deeply concerned about issues of *reproducibility* and *replicability* of scientific conclusions. Without getting into definitions and distinctions of these terms, we observe that much confusion and even doubt about the validity of science is arising. Such doubt can lead to radical choices, such as the one taken by the editors of *Basic and Applied Social Psychology*, who decided to ban *p*-values (null hypothesis significance testing) (Trafimow and Marks 2015). Misunderstanding or misuse of statistical inference is only one cause of the "reproducibility crisis" (Peng 2015), but to our community, it is an important one.

When the ASA Board decided to take up the challenge of developing a policy statement on *p*-values and statistical significance, it did so recognizing this was not a lightly taken step. The ASA has not previously taken positions on specific matters of statistical practice. The closest the association has come to this is a statement on the use of value-added models (VAM) for educational assessment (Morganstein and Wasserstein 2014) and a statement on risk-limiting post-election audits (American Statistical Association 2010). However, these were truly policy-related statements. The VAM statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of VAMs as effective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific policy issue (close elections in 2008), and said that statistically based election audits should become a routine part of election processes.

By contrast, the Board envisioned that the ASA statement on *p*-values and statistical significance would shed light on an aspect of our field that is too often misunderstood and misused in the broader research community, and, in the process, provides the community a service. The intended audience would be researchers, practitioners, and science writers who are not primarily statisticians. Thus, this statement would be quite different from anything previously attempted.

The Board tasked Wasserstein with assembling a group of experts representing a wide variety of points of view. On behalf of the Board, he reached out to more than two dozen such people, all of whom said they would be happy to be involved. Several expressed doubt about whether agreement could be reached, but those who did said, in effect, that if there was going to be a discussion, they wanted to be involved.

Over the course of many months, group members discussed what format the statement should take, tried to more concretely visualize the audience for the statement, and began to find points of agreement. That turned out to be relatively easy to do, but it was just as easy to find points of intense disagreement.

The time came for the group to sit down together to hash out these points, and so in October 2015, 20 members of the group met at the ASA Office in Alexandria, Virginia. The 2-day meeting was facilitated by Regina Nuzzo, and by the end of the meeting, a good set of points around which the statement could be built was developed.

The next 3 months saw multiple drafts of the statement, reviewed by group members, by Board members (in a lengthy discussion at the November 2015 ASA Board meeting), and by members of the target audience. Finally, on January 29, 2016, the Executive Committee of the ASA approved the statement.

The statement development process was lengthier and more controversial than anticipated. For example, there was considerable discussion about how best to address the issue of multiple *potential* comparisons (Gelman and Loken 2014). We debated at some length the issues behind the words "a *p*-value near 0.05 taken by itself offers only weak evidence against the null

© 2016 American Statistical Association

Wasserstein e Lazar (2016)

Apêndice
DEEP DIVE 1:
Por que estatística t apresenta distribuição t_{n-k-1}?

Não será cobrado

Ao ponderarmos $\beta_j$hat por seu erro padrão, obtemos uma estatística de teste que segue a distribuição t

- **se($\beta_j$hat)** é uma **variável aleatória**, pois depende de uma variável aleatória ($\beta_j$hat)

- A **razão** entre uma distribuição **normal padronizada** e a raiz quadrada de uma distribuição **qui-quadrado** dividida por seus graus de liberdade (como a distribuição de se($\beta_j$hat)) corresponde a uma **distribuição t**

$$\frac{\hat{\beta}_j - \beta_j^{Nula}}{EP(\hat{\beta}_j)} \sim t$$

Formalmente, se z *é uma variável aleatória com distribuição normal padrão e* x *é uma variável* χ^2 *[chi-quadrado] com* n *graus de liberdade, então a função abaixo comporta-se como uma distribuição t com* n *graus de liberdade:*

$$t(n) = \frac{z}{\sqrt{x/n}}$$

Bailey (2016: 785-6)

- $B_j$hat não é necessariamente z, mas a divisão de $B_j$hat pelo seu erro padrão é equivalente a uma z dividida pela raiz quadrada de (χ^2/graus de liberdade)

- Lembrando: uma distribuição $\mathbf{t_{GdeL}}$ **é semelhante a uma N(1, 0)**; porém, a t possui **caudas mais pesadas**

DEEP DIVE 1: Por que estatística t apresenta distribuição t_{n-k-1}? (1/3)

- Distribuição de $\beta_1$hats: compreende **todos os possíveis valores que $\beta_1$hat pode assumir**, e a **probabilidade** relativa desses valores

- Essa distribuição, chamada **distribuição amostral de $\beta_1$hat**, é uma distribuição **"teórica"**: não nos é visível, pois observamos apenas o $\beta_1$hat gerado por nossa (tipicamente única) amostra

- **Fórmula** do $\beta_1$hat pode ser **reescrita como uma média** (Bailey, 2016: 85, nota 8)

- Como tal, **$\beta_1$hat é uma estatística amostral sujeita ao Teorema Central do Limite** (aka Teorema do Limite Central), segundo o qual:

  – A média amostral de qualquer variável aleatória segue uma distribuição normal, com média = média da distribuição teórica (i.e., $\mu = \beta_1$)

  – Quanto maior o tamanho da amostra, mais próxima de uma normal será a distribuição amostral da média

Não será cobrado

DEEP DIVE 1:
Por que estatística t apresenta distribuição $t_{n\text{-}k\text{-}1}$? (2/3)

- Se o **erro estocástico (ε_i) for normalmente distribuído**, a **variância da distribuição teórica de $\beta_1$hat** será:

$$var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{n * var(X)} \qquad var(\hat{\beta}_1) \cong \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x - \bar{x})^2}$$

onde σ^2 = variância de ε, que é desconhecida;
σ^2 é estimada por σ^2hat, a variância da regressão, que corresponde à soma dos quadrados dos resíduos dividida pelos graus de liberdade do modelo (n – k – 1):

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - k - 1} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{n - k - 1}$$

- Quando apuramos a distância entre a evidência amostral ($\beta_1$hat) e o cenário da hipótese nula (β_1^{Nula}), e ponderamos essa distância pela imprecisão da evidência amostral (i.e., pelo erro padrão de $\beta_1$hat, o se($\beta_1$hat)), estamos calculando a seguinte razão:

$$\frac{\hat{\beta}_j - \beta_j^{Nula}}{EP(\hat{\beta}_j)}$$

- Que, na **regressão simples e na regressão múltipla isenta de multicolinearidade**, pode ser reescrita como a razão entre uma distribuição normal padrão e a raiz quadrada de uma distribuição qui-quadrado dividida por seus graus de liberdade (vide próxima página)

- Ou seja, a estatística t se comporta como uma distribuição t com n – k – 1 graus de liberdade

Não será cobrado

DEEP DIVE 1:
Por que estatística t apresenta distribuição t_{n-k-1}? (3/3)

$$\frac{\hat\beta_1 - \beta_1^{Nula}}{\sqrt{\dfrac{\hat\sigma^2}{\cancel{N}\cdot\dfrac{\sum(X-\bar X)^2}{\cancel{N-1}}}}} \overset{*}{=} \frac{\boxed{\dfrac{\hat\beta_1-\beta_1^{Nula}}{\sqrt{\sigma^2(\hat\beta_1)}}}}{\dfrac{\sqrt{\dfrac{\hat\sigma^2}{\sum(X-\bar X)^2}}}{\sqrt{\sigma^2(\beta_1)}}} = \frac{z}{\sqrt{\dfrac{\dfrac{\hat\sigma^2}{\cancel{\sum(X-\bar X)^2}}}{\boxed{\dfrac{\sigma^2}{\cancel{\sum(X-\bar X)^2}}}}}}$$

numerador agora corresponde a uma distribuição normal padrão *

pode ser substituído por N se amostra for grande

Apenas dividimos numerador e denominador por $\sqrt{\sigma^2(\hat\beta_1)}$, ou seja, pelo desvio padrão da distribuição teórica de $\hat\beta_1$; esse desvio padrão é desconhecido

Para uma regressão com erros normalmente distribuídos, $\sigma^2(\hat\beta_1)$ é dada por $\frac{\sigma^2}{\sum(X-\bar X)^2}$, em que σ^2 é a variância de ε

$$= \frac{z}{\sqrt{\dfrac{\dfrac{\sum(Y-\hat Y)^2}{n-k-1}}{\dfrac{\sigma^2}{1}}}} = \frac{z}{\sqrt{\boxed{\dfrac{\sum(Y-\hat Y)^2}{\sigma^2}}\cdot\dfrac{1}{n-k-1}}} = \frac{z}{\sqrt{\dfrac{\chi^2_{n-k-1}}{n-k-1}}} \sim t_{n-k-1}$$

Pelo Teorema de Cochran, a soma dos quadrados dos resíduos dividida pela variância do ε corresponde a uma distribuição χ^2 com $n-k-1$ graus de liberdade

* $z = \dfrac{X_i - \mu}{\sigma(x)}$, ou seja, distribuição z, que é normal padrão, corresponde à diferença entre cada X_i e a média de X, dividida pelo desvio-padrão de X.

Não será cobrado

Apêndice
DEEP DIVE 2:
Qual a intuição por trás do intervalo de confiança?

Não será cobrado

DEEP DIVE 2:
Qual a intuição por trás do intervalo de confiança? (1/4)

Dada a estimativa βhat que obtivemos, quais valores de β seriam consistentes com essa estimativa (i.e., “prováveis”)? Para responder, apoiamo-nos nas afirmações abaixo.

A **Na ausência de endogeneidade, E(βhat) = β**
Se não houver viés, β é o centro da distribuição de βhat e o valor mais provável de βhat

Implicação: Não sabemos quão próxima de β nossa estimativa se encontra, mas podemos acreditar que provavelmente está próxima, a depender do grau de precisão

B **Distribuição teórica das estimativas de βhat é simétrica**
Se βhat for diferente de β, há igual probabilidade de que βhat esteja à direita ou à esquerda de β

Implicação: Podemos estimar uma banda de variação para β centrada em βhat – até porque não teríamos opção melhor para “ancorar” esse intervalo

C **Aleatoriedade na estimação de βhat advém tanto de sampling randomness quanto de modeled randomness**
Na ausência de endogeneidade, as fontes de incerteza sobre βhat são a amostragem (sampling randomness) e a escolha de fatores considerados no modelo explicativo (modeled randomness)

Implicação: Intervalo de confiança deve considerar a aleatoriedade advinda tanto da sampling randomness quanto da modeled randomness

Não será cobrado

DEEP DIVE 2:
Qual a intuição por trás do intervalo de confiança? (2/4)

Dada a estimativa βhat que obtivemos, quais valores de β seriam consistentes com essa estimativa (i.e., "prováveis")? Para responder, apoiamo-nos nas afirmações abaixo.

D **t crítico é uma medida de sampling randomness**

Dados os graus de liberdade, t crítico informa uma distância em relação a β^{Nula}; a probabilidade de valores de t ≥ t crítico (ou t ≤ -t crítico) é de (1 – nível de confiança)/2

Implicações:

- t crítico não é influenciado pela escolha de variáveis explicativas, apenas:
  - Pelo nível de confiança considerado; e
  - Pelos graus de liberdade: n - k - 1 (i.e., número de observações livres)

- Portanto, o t crítico reflete apenas um atributo preestabelecido da estimativa intervalar (nível de confiança) e apenas um atributo do modelo (número de graus de liberdade, diretamente afetado pelo tamanho da amostra)
  Nota: o t crítico é o mesmo para qualquer βhat de um modelo

- t crítico pode ser usado para representar a sampling randomness (quanto maior a amostra, menor a sampling randomness e maiores os graus de liberdade)

Não será cobrado

DEEP DIVE 2:
Qual a intuição por trás do intervalo de confiança? (3/4)

Dada a estimativa βhat que obtivemos, quais valores de β seriam consistentes com essa estimativa (i.e., "prováveis")? Para responder, apoiamo-nos nas afirmações abaixo.

E **Erro padrão é uma medida de sampling e modeled randomness combinadas**

O erro padrão informa o grau de imprecisão da distribuição teórica de βhat; essa imprecisão é causada por sampling e modeled randomness

Implicações:

- O erro padrão é influenciado:
  - Pelo tamanho da amostra: no denominador (no N e variância de X_j)
  - Pelos graus de liberdade: no numerador (na fórmula da variância da regressão);
  - Pela escolha de variáveis explicativas: no numerador (na variância da regressão, que é uma medida de ajuste – quanto maior a variância da regressão, menor o ajuste);
  - Pela variação de X_j: no denominador (na variância de X_j); e
  - [Na regressão múltipla] Pela independência entre variáveis explicativas: no denominador (em 1 – R_j^2)

- Portanto, o erro padrão reflete os dois tipos de aleatoriedade (tanto sampling quanto modeled randomness)
  Nota: O erro padrão é específico ao βhat para o qual é estimado

- Erro padrão pode ser usado para representar sampling e modeled randomness

Não será cobrado

DEEP DIVE 2:
Qual a intuição por trás do intervalo de confiança? (4/4)

Dada a estimativa βhat que obtivemos, quais valores de β seriam consistentes com essa estimativa (i.e., “prováveis”)? Para responder, apoiamo-nos nas afirmações abaixo.

F **Para corresponder a um certo nível de confiança, intervalo de confiança deve combinar t crítico e erro padrão**

Considerando o nível de confiança e os graus de liberdade, t crítico informa sobre a região onde β é provável de estar (dado o βhat estimado), mas é "genérico": para um certo nível de confiança, será o mesmo para qualquer modelo com os mesmos graus de liberdade

A seu turno, erro padrão reflete sampling e modeled randomness, mas nada diz sobre a região onde β é mais provável de estar (dado o βhat estimado)

Implicações:

- Intervalo de confiança não pode se apoiar unicamente no t crítico ou no erro padrão, devendo combinar essas medidas de aleatoriedade

- A um certo nível de confiança, ao se multiplicar t crítico pelo erro padrão:
  - Se erro padrão for igual à unidade, o intervalo de confiança é determinado pelo t crítico
  - Se erro padrão for inferior à unidade, estreita-se o intervalo de confiança
  - Se erro padrão for superior à unidade, dilata-se o intervalo de confiança

Não será cobrado

ANEXO

Ernesto Amaral

Não seguro | ernestoamaral.com/dcp046-141.html

Home

ERNESTO F. L. AMARAL

Departamento de Ciência Política (DCP)
Faculdade de Filosofia e Ciências Humanas (FAFICH)
Universidade Federal de Minas Gerais (UFMG)

(2014-1b) DCP 046 – Avaliação de Políticas Públicas
Período: 1º semestre de 2014
Horário: terças-feiras (20:50 às 22:30) e sextas-feiras (19:00 às 20:40)
Aulas teóricas: FAFICH 3011
Aulas práticas: FAFICH 3062
Atendimento aos alunos: FAFICH 209, terças-feiras (14:00 às 19:00), com marcação por email
Carga horária: 60 horas/aula (4 créditos)

Professor:
Ernesto Friedrich de Lima Amaral
Email: eflamaral@gmail.com

http://www.ernestoamaral.com/dcp046-141.html

Tipos de teste t: Slides do Prof. Ernesto Amaral

(DCP134 – Avaliação de Políticas Públicas, 2014a, Aula 20-21)

8

TESTE: HIPÓTESES ALTERNATIVAS UNILATERAIS

$$H_1: \beta_j > 0 \quad \text{OU} \quad H_1: \beta_j < 0$$

– Devemos decidir sobre um nível de significância (geralmente de 5%).

– Corremos o risco de rejeitar erroneamente H_0, quando ela é verdadeira, em 5% das vezes (erro tipo I igual ao α).

– Um valor suficientemente grande de *t*, com um nível de significância de 5%, é o 95º percentil de uma distribuição *t* com n-k-1 graus de liberdade (ponto c).

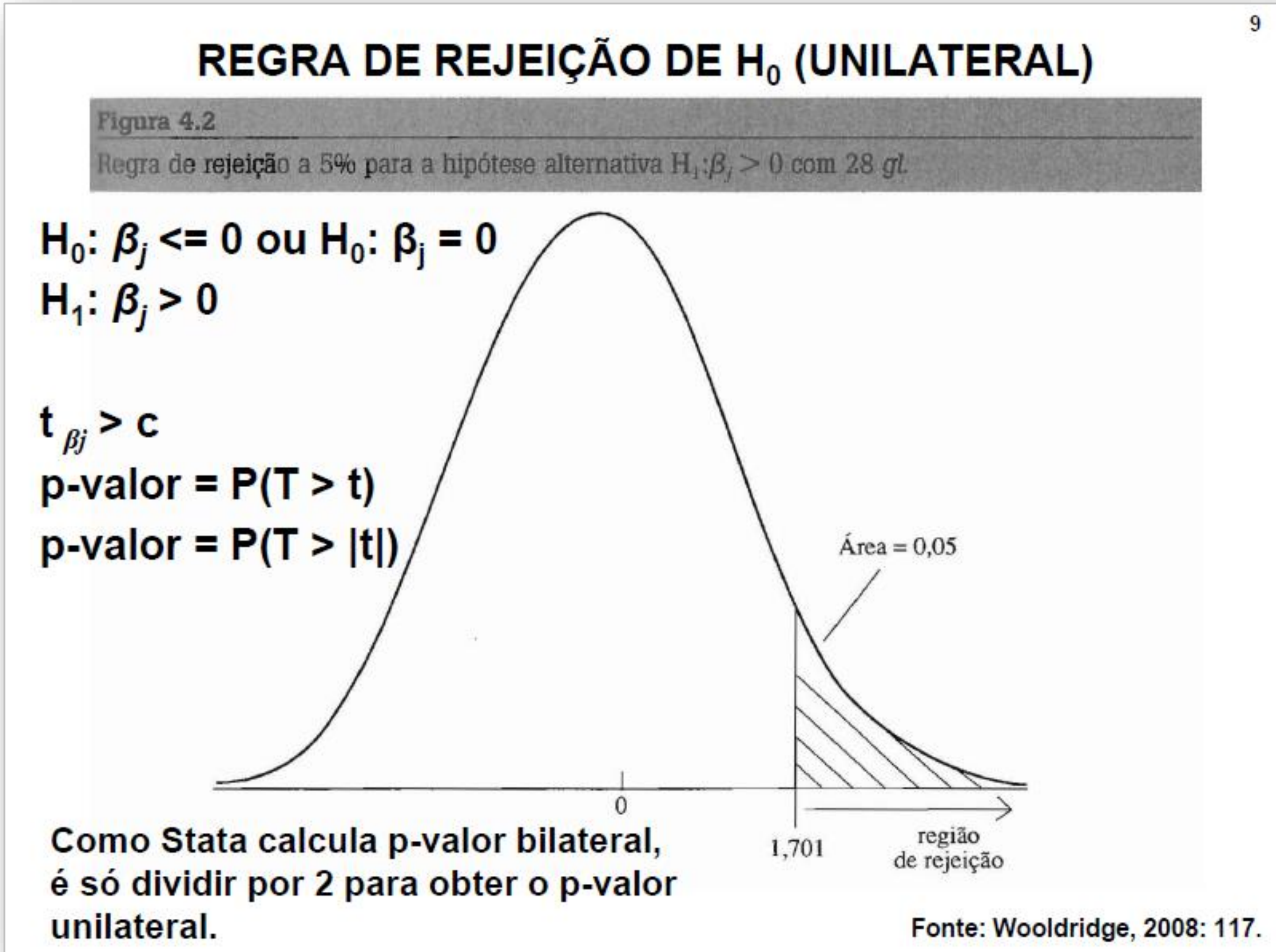
– **Regra de rejeição** é que H_0 é rejeitada em favor de H_1, se $t>c$ (H_1:$\beta_j>0$) ou $t<-c$ (H_1:$\beta_j<0$), em um nível específico.

– Quando os graus de liberdade da distribuição *t* ficam maiores, a distribuição *t* aproxima-se da distribuição normal padronizada.

– Para graus de liberdade maiores que 120, pode-se usar os valores críticos da distribuição normal padronizada.

Tipos de teste t: Slides do Prof. Ernesto Amaral

(DCP134 – Avaliação de Políticas Públicas, 2014a, Aula 20-21)



Tipos de teste t: Slides do Prof. Ernesto Amaral

(DCP134 – Avaliação de Políticas Públicas, 2014a, Aula 20-21)

10

REGRA DE REJEIÇÃO DE H_0 (UNILATERAL)

Figura 4.3

Regra de rejeição a 5% para a hipótese alternativa H_1: $\beta_j < 0$, com 18 *gl*.

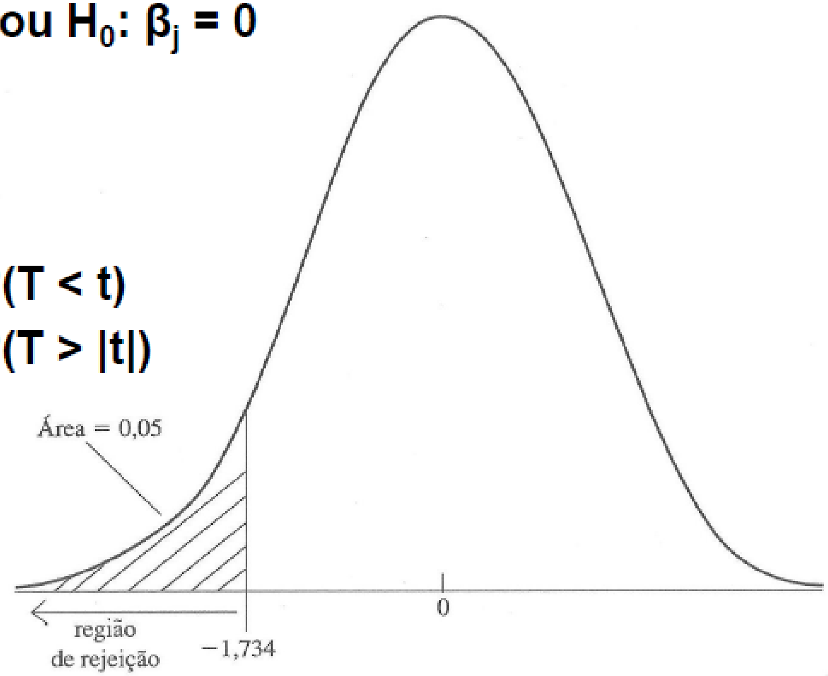
H_0: $\beta_j >= 0$ ou H_0: $\beta_j = 0$

H_1: $\beta_j < 0$

$t_{\beta j} < -c$

p-valor = P(T < t)

p-valor = P(T > |t|)



Como Stata calcula p-valor bilateral, é só dividir por 2 para obter o p-valor unilateral.

Fonte: Wooldridge, 2008: 119.

Tipos de teste t: Slides do Prof. Ernesto Amaral

(DCP134 – Avaliação de Políticas Públicas, 2014a, Aula 20-21)

11

TESTE: HIPÓTESES ALTERNATIVAS BILATERAIS

H_1: $\beta_j \neq 0$

– Essa hipótese é relevante quando o sinal de β_j não é bem determinado pela teoria.

– Usar as estimativas da regressão para nos ajudar a formular as hipóteses nula e alternativa não é permitido, porque a inferência estatística clássica pressupõe que formulamos as hipóteses nula e alternativa sobre a população antes de olhar os dados.

– Quando a alternativa é bilateral, estamos interessados no valor absoluto da estatística *t*: $|t|>c$.

– Para um nível de significância de 5% e em um teste bi-caudal, *c* é escolhido de forma que a área em cada cauda da distribuição *t* seja igual a 2,5%.

– Se H_0 é rejeitada, x_j é estatisticamente significante (ou estatisticamente diferente de zero) ao nível de 5%.

Tipos de teste t: Slides do Prof. Ernesto Amaral

(DCP134 – Avaliação de Políticas Públicas, 2014a, Aula 20-21)

REGRA DE REJEIÇÃO DE H_0 (BILATERAL)

Figura 4.4

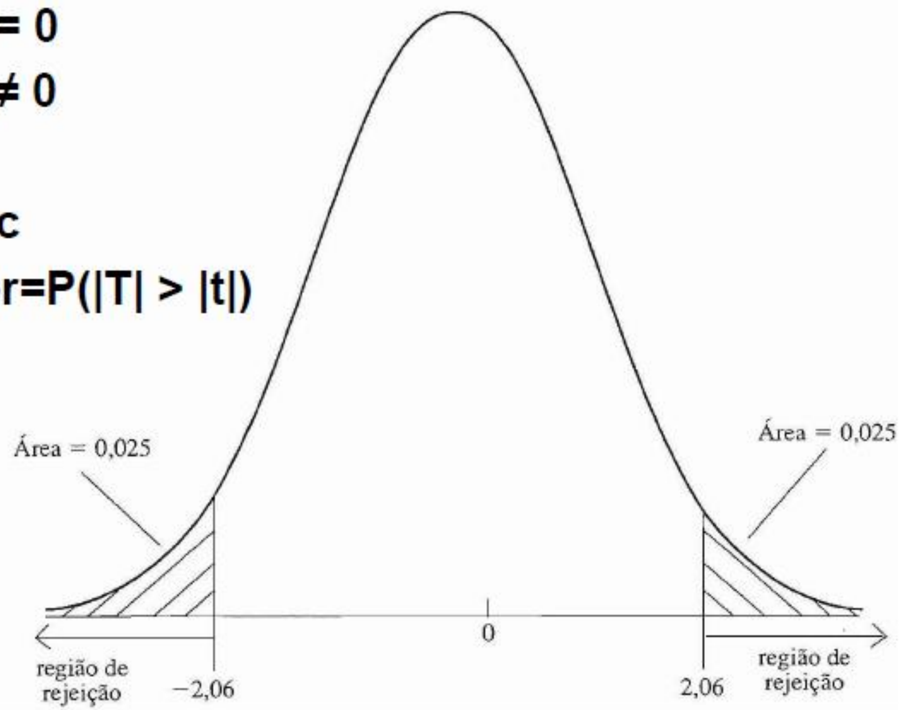
Regra de rejeição a 5% para a hipótese alternativa $H_1: \beta_j \neq 0$ com 25 *gl*.

$H_0: \beta_j = 0$

$H_1: \beta_j \neq 0$

$|t_{\beta j}| > c$

p-valor=$P(|T| > |t|)$



Fonte: Wooldridge, 2008: 122.

Tipos de teste t: Slides do Prof. Ernesto Amaral

(DCP134 – Avaliação de Políticas Públicas, 2014a, Aula 20-21)

13

EXEMPLO DE NÃO-REJEIÇÃO DE H_0 (BILATERAL)

Figura 4.6

Obtendo o *p*-valor contra uma alternativa bilateral, quando $t = 1{,}85$ e $gl = 40$.

p-valor

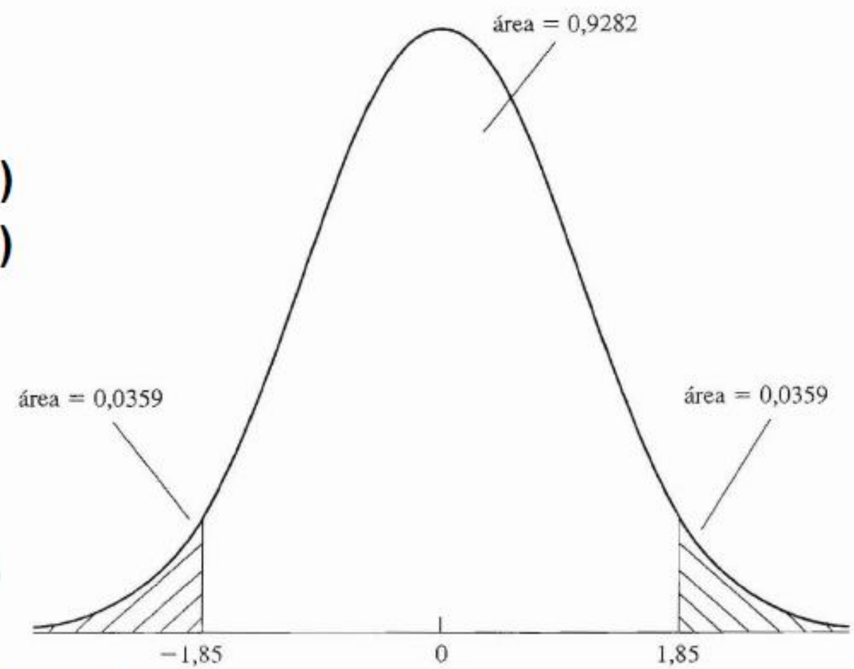
= P(|T| > |t|)

= P(|T| > 1,85)

= 2P(T > 1,85)

= 2(0,0359)

= 0,0718



p-valor > α

0,0718 > 0,05

H_0 : β_j=0 não é rejeitada

Fonte: Wooldridge, 2008: 127.

Tipos de teste t: Slides do Prof. Ernesto Amaral

(DCP134 – Avaliação de Políticas Públicas, 2014a, Aula 20-21)

TESTES DE OUTRAS HIPÓTESES SOBRE β_j

– Poderíamos supor que uma variável dependente (log do número de crimes) necessariamente será relacionada positivamente com uma variável independente (log do número de estudantes matriculados na universidade).

– A hipótese alternativa testará se o aumento de 1% nas matrículas aumentará o crime em mais de 1%:

$$H_0: \beta_j = 1$$

$$H_1: \beta_j > 1$$

– t = (estimativa - valor hipotético) / (erro-padrão)

– Neste exemplo, $t = (\beta_j - 1) / ep(\beta_j)$

– Observe que adicionar 1 na hipótese nula, significa subtrair 1 no teste *t*.

– Rejeitamos H_0 se $t > c$, em que c é o valor crítico unilateral.

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA POLÍTICA | UFMG

DCP919B
Metodologia II

Teste de hipótese

Aula 4

28 de setembro de 2022

Ana Paula Karruz