



escola  
britânica de  
artes criativas  
& tecnologia

## **Profissão: Cientista de Dados**

Análise de Regressão II

# Previsão vs Explicação



**Sir Ronald Fisher**

# Previsão vs explicação



**Alan Turing**

# Previsão vs explicação

## Previsão:

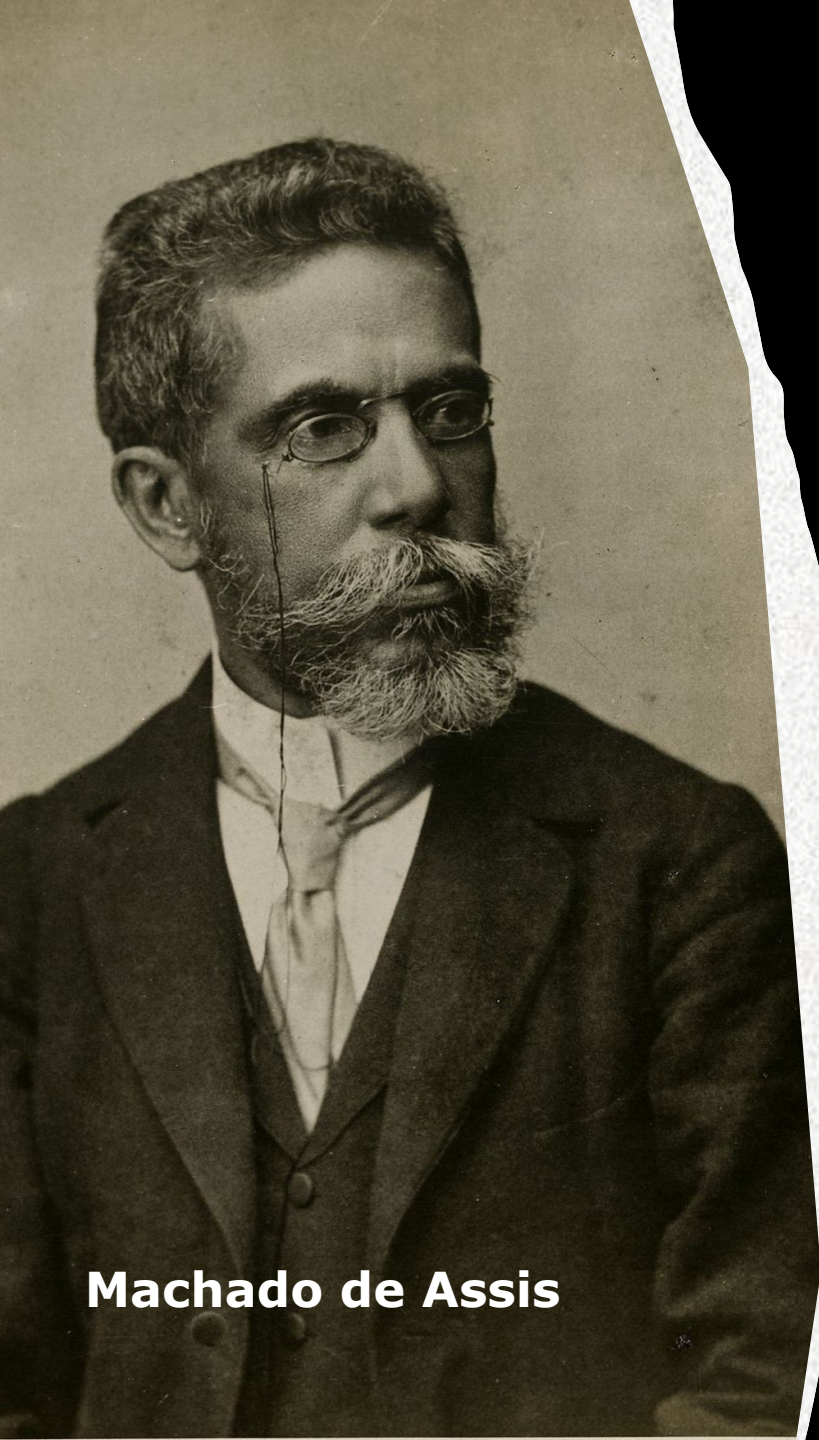
- Objetivo mais comum em *machine learning*
- Encontrar padrões em dados e “prever” novas observações
- Poucas premissas
- Foco em redução de erros

## Explicação:

- Objetivo mais comum em estatística
- Identificar associações e interpretá-las
- Testar hipóteses sobre elas
- Premissas mais fortes sobre os dados
- Mede erro amostral

Para descontraír...





**Machado de Assis**

Como se  
pareciam...



**Ronald Fisher**

# Inferência sobre os parâmetros

# Premissas sobre os dados

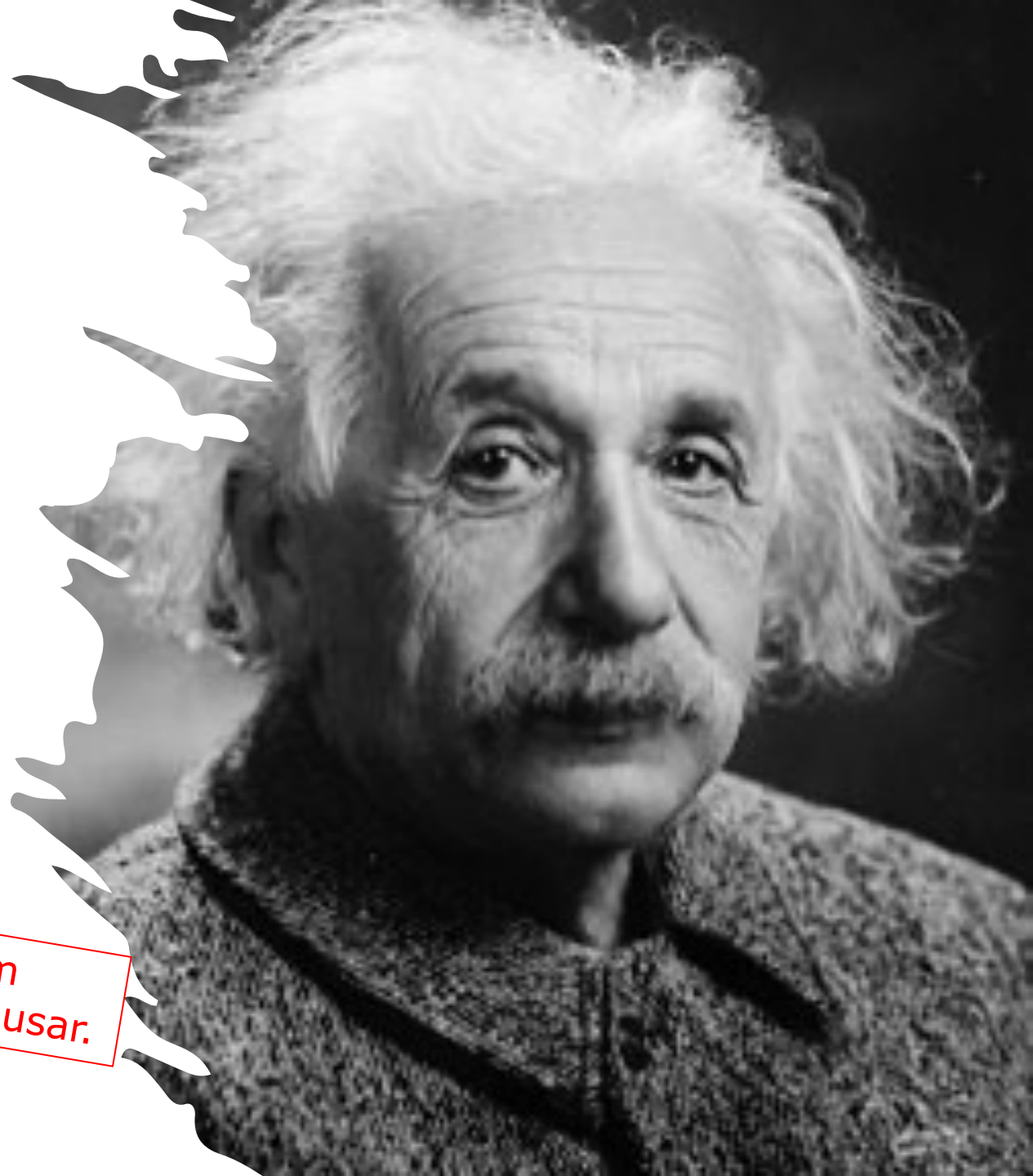
- São observações aleatórias seja devido a:
  - Processo de amostragem
  - Processo estocástico de geração
  - Por premissa
- Possuem alguma “lei” geradora sobre a qual queremos inferir



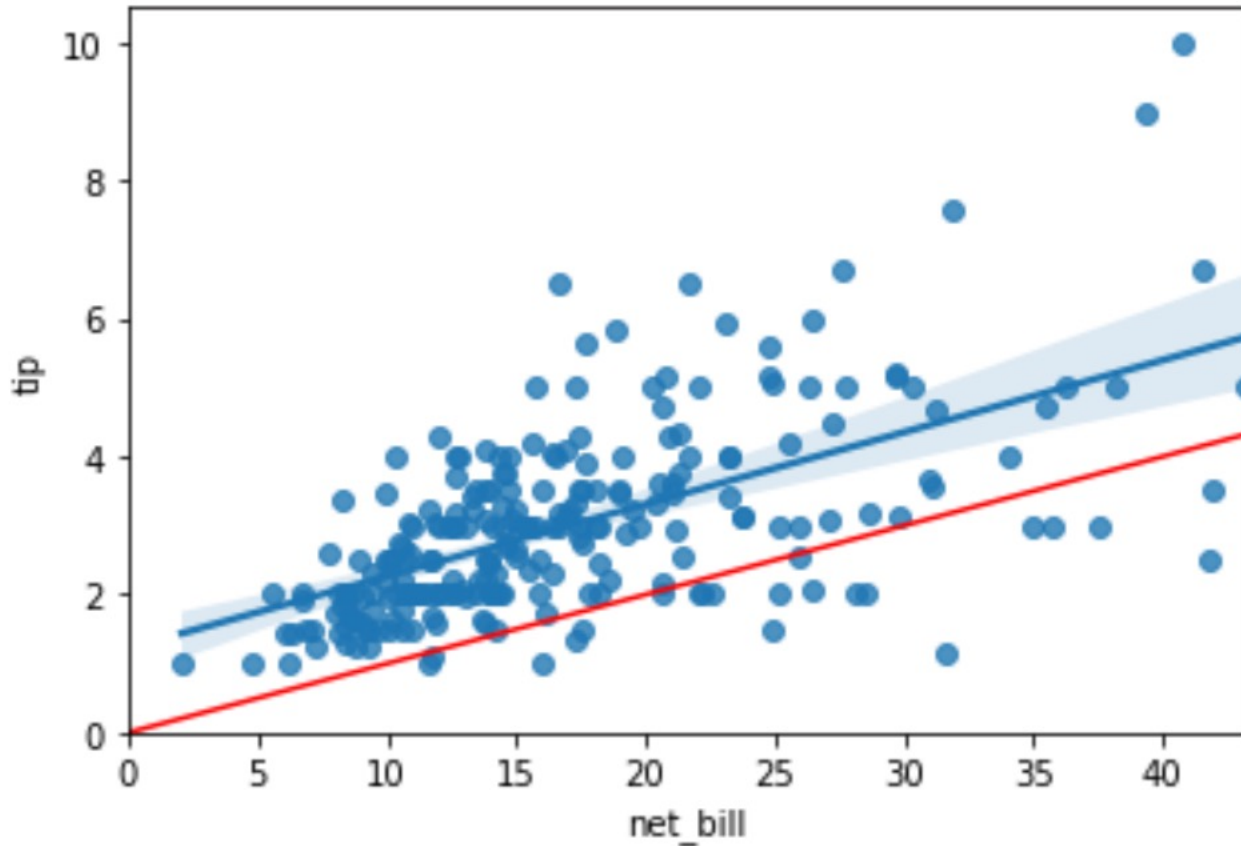
"As I have said so many times, God  
doesn't play dice with the world."

Albert Einstein

Einstein não estava errado... Só trabalhava com um  
modelo que não é útil sob o paradigma que vamos usar.



# Uma estimativa é aleatória

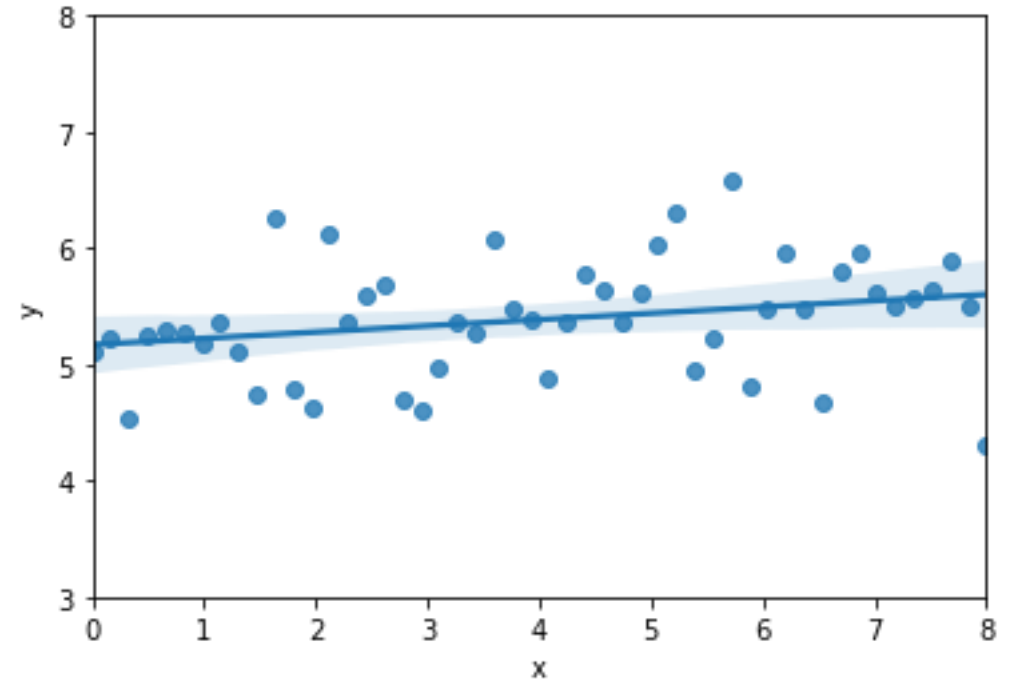
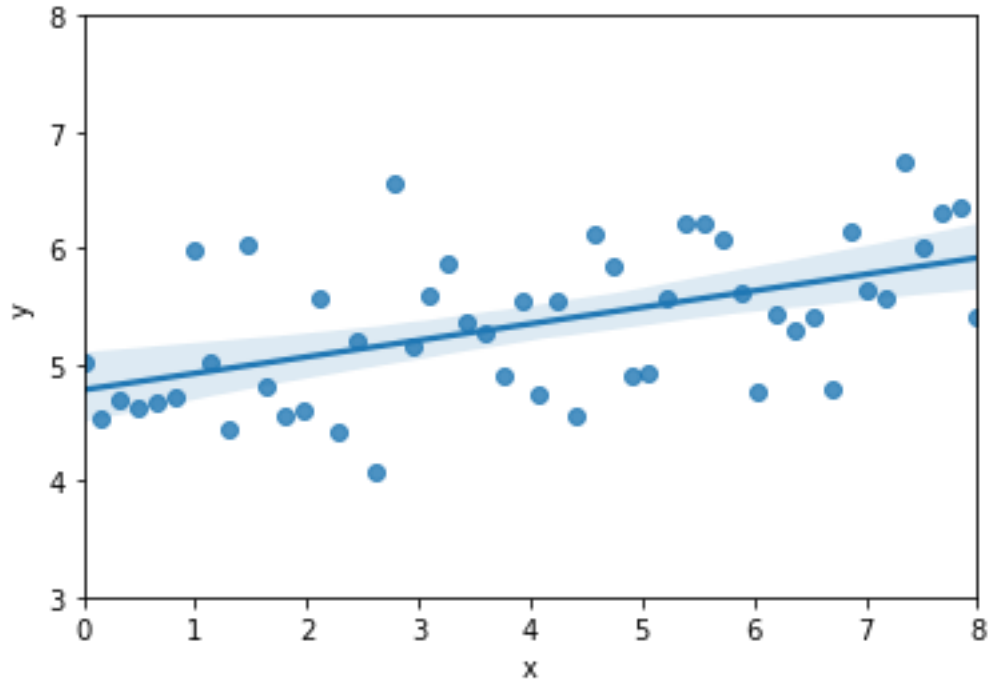


Siga o meu raciocínio:

- Vemos dados são variáveis aleatórias
- $\hat{\beta}$  é uma função dos dados (x e y)
- $\hat{\beta}$  é, portanto, variável aleatória também.
- Se conhecermos a 'lei' de  $\hat{\beta}$  podemos fazer inferência sobre ele

Qual será a "lei" geradora de  $\hat{\beta}$ ?

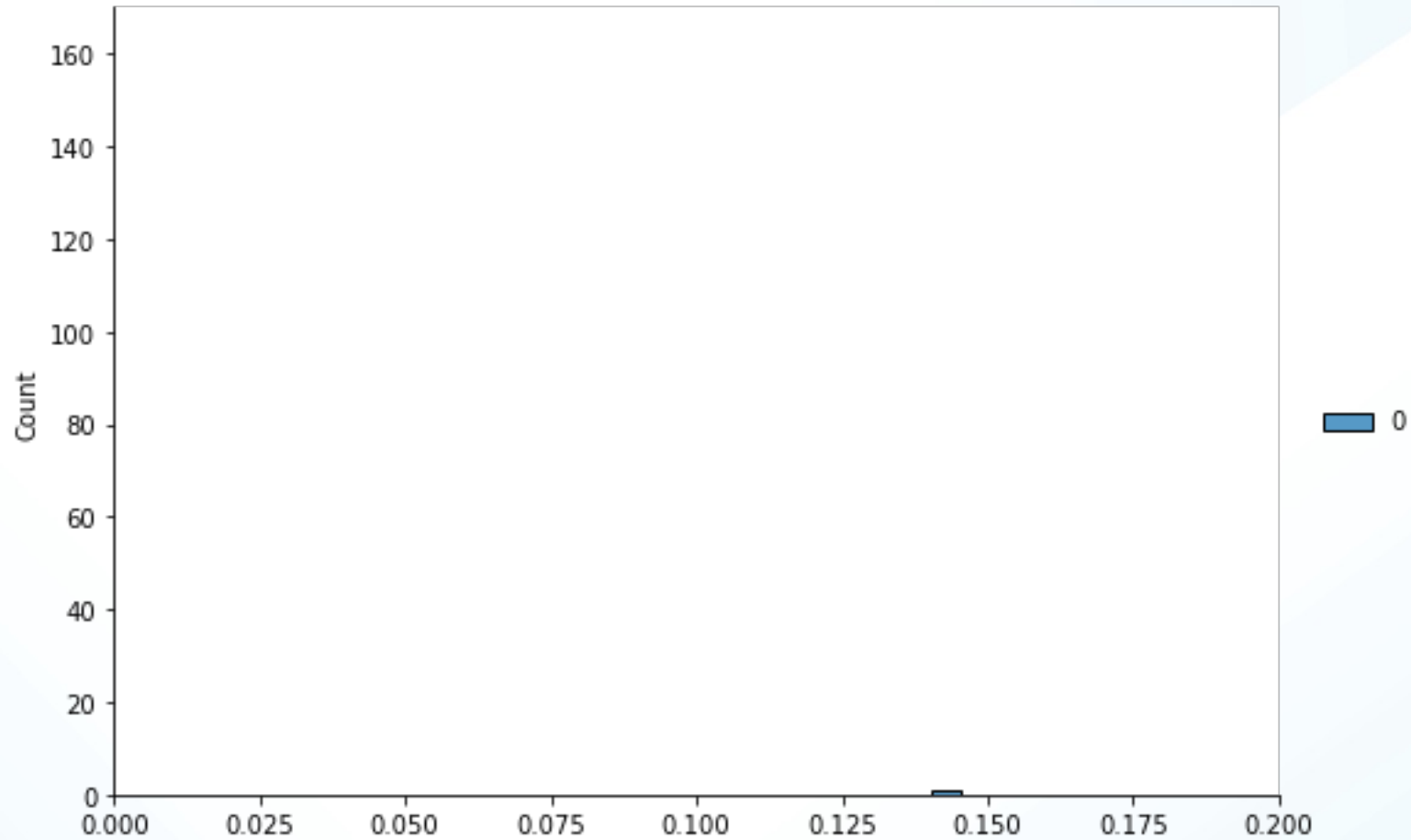
# Raciocínio com dados simulados



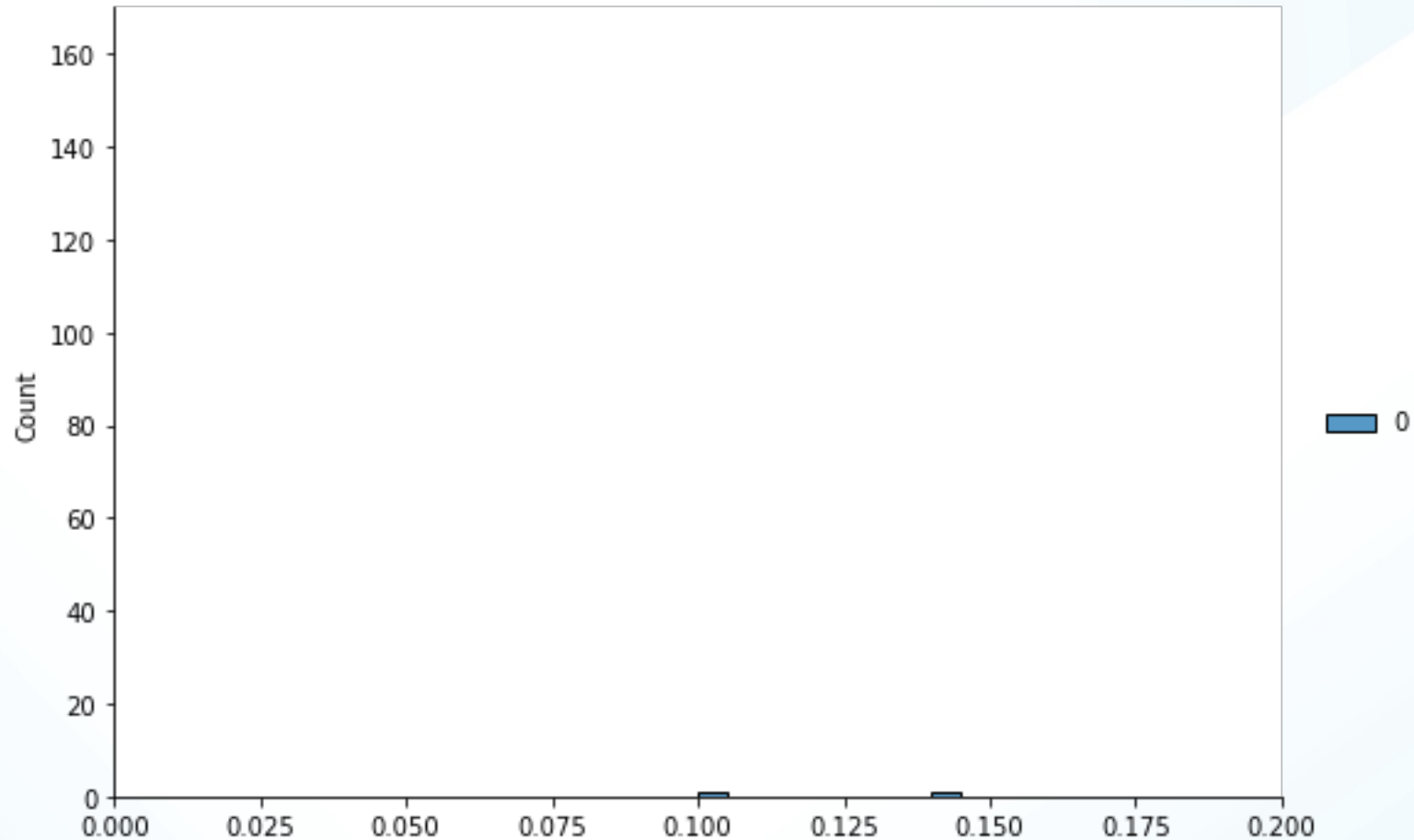
$$y = 5 + 0.1x + \epsilon$$

com o parâmetro aleatório de erro sendo:  $\epsilon \sim N(0, 0.25)$

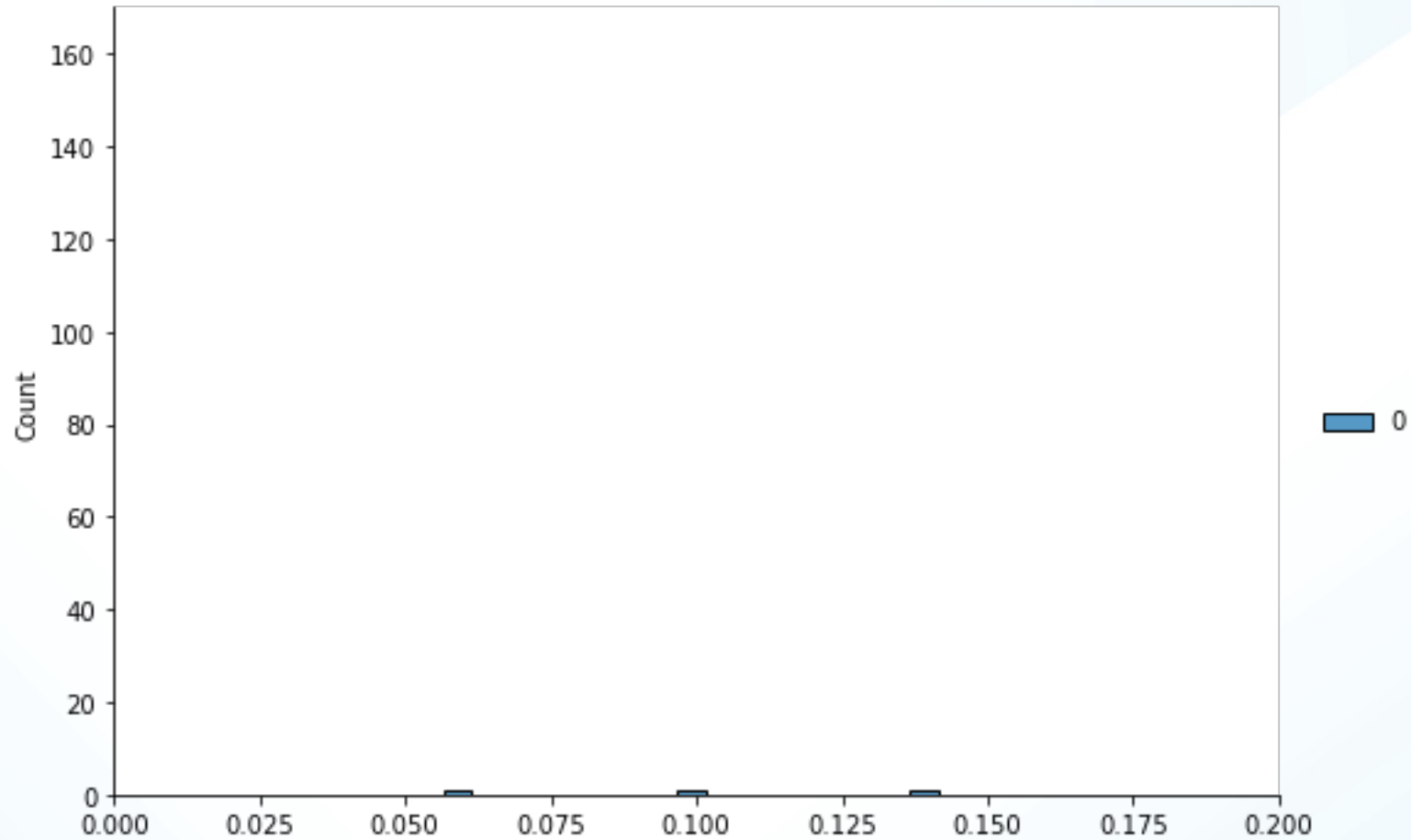
# Como é a “lei” geradora do beta?



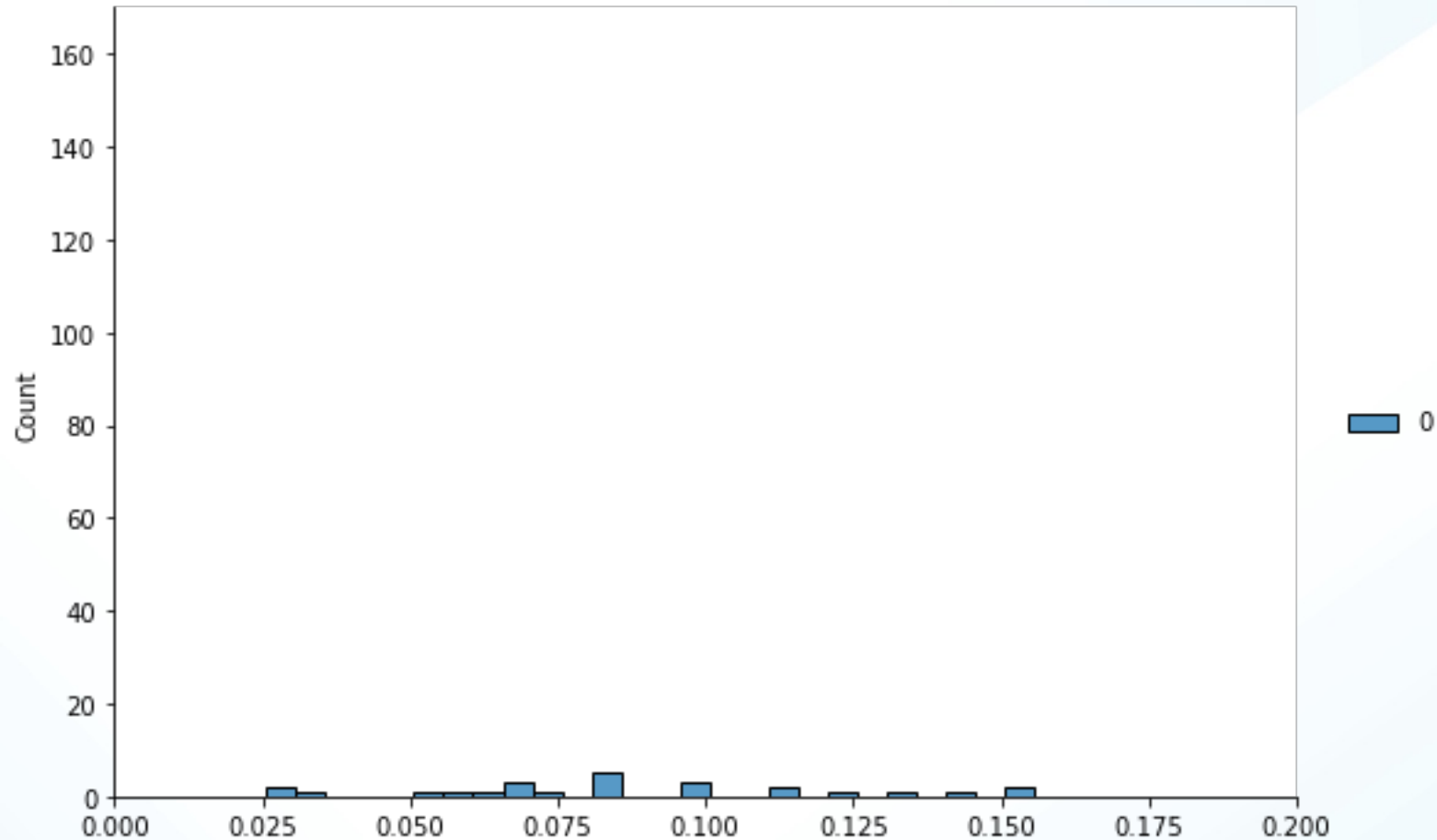
# Como é a “lei” geradora do beta?



# Como é a “lei” geradora do beta?

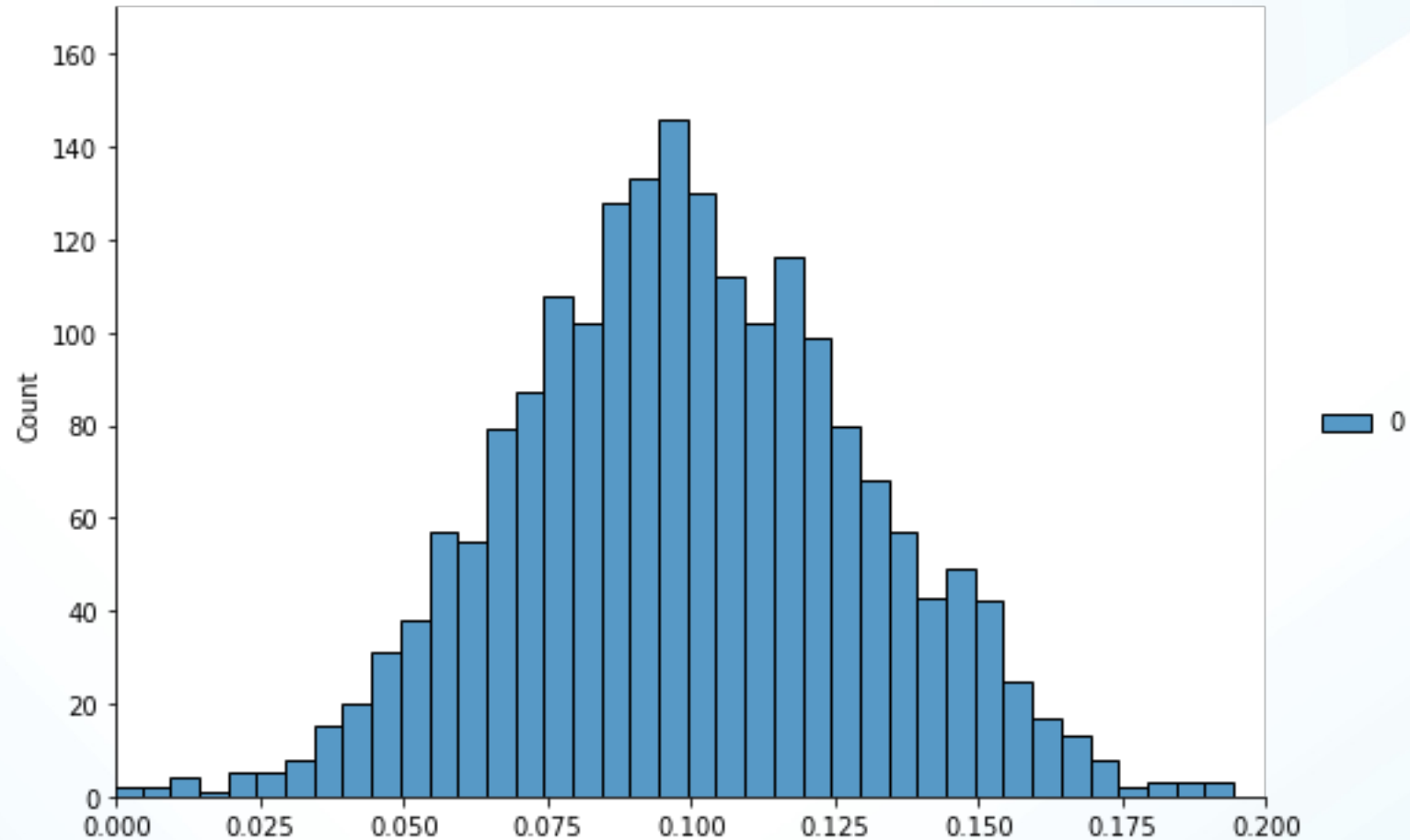


# Como é a “lei” geradora do beta?





# Como é a “lei” geradora do beta?



# Premissas da regressão

- Erros  $\sim N(0, \sigma^2)$
- Erros independentes
- Homocedasticidade (variância homogênea)

# Estabelecendo a hipótese a se testar

- Formulando a hipótese nula:

$$H_0: \beta = 0$$

$$H_a: \beta \neq 0$$

$$t = \frac{\hat{\beta}}{\text{erro} - \text{padrão}_{\hat{\beta}}} \sim N(\beta, \sigma^2_{\beta})$$

	coef	std err	t	P> t	[0.025	0.975]
<b>Intercept</b>	5.1700	0.136	37.988	0.000	4.896	5.444
<b>x</b>	0.0540	0.029	1.842	0.072	-0.005	0.113

# Estabelecendo a hipótese a se testar

- Formulando a hipótese nula:

$$H_0: \beta = 0$$

$$H_a: \beta \neq 0$$

$$t = \frac{\hat{\beta}}{\text{erro} - \text{padrão}_{\hat{\beta}}} \sim N(\beta, \sigma^2_{\beta})$$

	coef	std err	t	P> t	[0.025	0.975]
<b>Intercept</b>	5.1700	0.136	37.988	0.000	4.896	5.444
<b>x</b>	0.0540	0.029	1.842	0.072	-0.005	0.113

# Estabelecendo a hipótese a se testar

- Formulando a hipótese nula:

$$H_0: \beta = 0$$

$$H_a: \beta \neq 0$$

$$t = \frac{\hat{\beta}}{\text{erro} - \text{padrão}_{\hat{\beta}}} \sim N(\beta, \sigma^2_{\beta})$$

$\hat{\beta}$



	coef	std err	t	P> t	[0.025	0.975]
<b>Intercept</b>	5.1700	0.136	37.988	0.000	4.896	5.444
<b>x</b>	0.0540	0.029	1.842	0.072	-0.005	0.113

Estimativa do beta

# Estabelecendo a hipótese a se testar

- Formulando a hipótese nula:

$$H_0: \beta = 0$$

$$H_a: \beta \neq 0$$

$$t = \frac{\hat{\beta}}{\text{erro} - \text{padrão}_{\hat{\beta}}} \sim N(\beta, \sigma^2_{\beta})$$

*erro - padrão <sub>$\hat{\beta}$</sub>*

$\hat{\beta}$

	coef	std err	t	P> t	[0.025	0.975]
<b>Intercept</b>	5.1700	0.136	37.988	0.000	4.896	5.444
<b>x</b>	0.0540	0.029	1.842	0.072	-0.005	0.113

Erro padrão do beta

# Estabelecendo a hipótese a se testar

- Formulando a hipótese nula:

$$H_0: \beta = 0$$

$$H_a: \beta \neq 0$$

$$t = \frac{\hat{\beta}}{\text{erro} - \text{padrão}_{\hat{\beta}}} \sim N(\beta, \sigma^2_{\beta})$$

*erro – padrão* $\hat{\beta}$

$\hat{\beta}$

*t*

	coef	std err	t	P> t	[0.025	0.975]
<b>Intercept</b>	5.1700	0.136	37.988	0.000	4.896	5.444
<b>x</b>	0.0540	0.029	1.842	0.072	-0.005	0.113

Estatística t – estimativa/erro padrão



# Estabelecendo a hipótese a se testar

- Formulando a hipótese nula:

$$H_0: \beta = 0$$

$$H_a: \beta \neq 0$$

$$t = \frac{\hat{\beta}}{\text{erro} - \text{padrão}_{\hat{\beta}}} \sim N(\beta, \sigma^2_{\beta})$$

*erro - padrão $\hat{\beta}$*  *t* *p - value*

$\hat{\beta}$

	coef	std err	t	P> t	[0.025	0.975]
Intercept	5.1700	0.136	37.988	0.000	4.896	5.444
x	0.0540	0.029	1.842	0.072	-0.005	0.113

P-value: probabilidade de observarmos um novo  $\hat{\beta}$  sob  $H_0$ , menos provável que a estimativa atual.

# Estabelecendo a hipótese a se testar

- Formulando a hipótese nula:

$$H_0: \beta = 0$$

$$H_a: \beta \neq 0$$

$$t = \frac{\hat{\beta}}{\text{erro} - \text{padrão}_{\hat{\beta}}} \sim N(\beta, \sigma^2_{\beta})$$

*erro - padrão $\hat{\beta}$*  *t* *p - value*  
 *$\hat{\beta}$*  *Intervalo de confiança*

	coef	std err	t	P> t	[0.025	0.975]
Intercept	5.1700	0.136	37.988	0.000	4.896	5.444
x	0.0540	0.029	1.842	0.072	-0.005	0.113

Podemos dizer, com 95% de confiança, que o verdadeiro valor do parâmetro está dentro desse intervalo.