Aula de hoje

Aula sobre

Introdução à Inferência Estatística

Distribuição da Média Amostral

Uso do TLC

1º: uso do notebook Aula18_TLC_interativo-vf

2º: uso do notebook Aula18_Atividade

3º: uso do notebook Aula18_Exercício

Ciência dos dados

Teorema do Limite Central - TLC

Insper

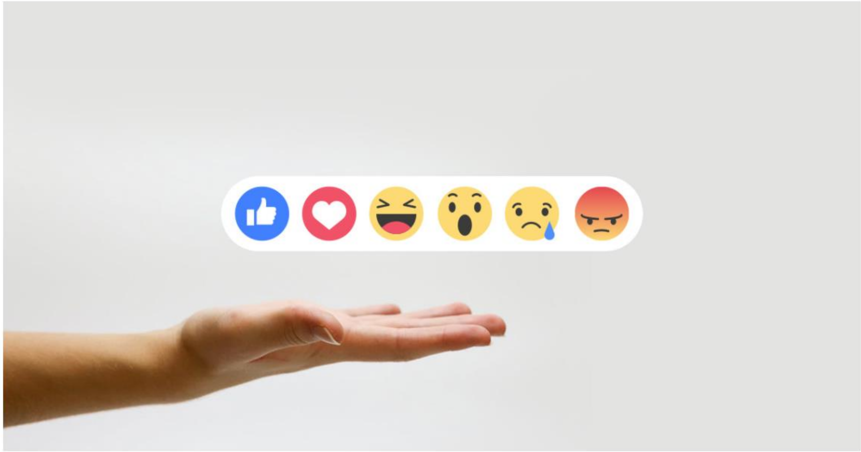
Introdução

Inferência Estatística é a área da estatística que contém métodos usados para tomar decisões usando uma amostra e tirar conclusões acerca de *populações*.

Motivação – Satisfação de clientes

PESQUISA DE SATISFAÇÃO DO CLIENTE: COMO ELA PODE AJUDAR SUA LOJA DE ROUPAS

9 DE JULHO DE 2018 | DICAS, LOJISTAS



"Como empreendedor, é do seu interesse que sua **loja de roupas** permaneça em constante desenvolvimento. Isso significa atrair e reter os clientes, qualificar a sua equipe e vender mais."

https://www.brixjeans.com.br/blog/pesquisa-de-satisfacao-do-cliente/

Motivação – Satisfação de clientes

Uma rede de lojas do setor de vestuário, com lojas nos principais shoppings do Brasil, com centenas de milhares de clientes, deseja avaliar a satisfação média de seus clientes.

Você foi contratado para avaliar essa satisfação. Como você faria isso?

Satisfação

Uma rede de lojas do setor de vestuário, com lojas nos principais shoppings do Brasil, com centenas de milhares de clientes, deseja avaliar a satisfação média de seus clientes.

Você foi contratado para avaliar essa satisfação. Como você faria isso?

Qual a variável de interesse?

X : *nível de satisfação, de cada cliente da rede de lojas*

Satisfação

Uma rede de lojas do setor de vestuário, com lojas nos principais shoppings do Brasil, com centenas de milhares de clientes, deseja avaliar a satisfação média de seus clientes.

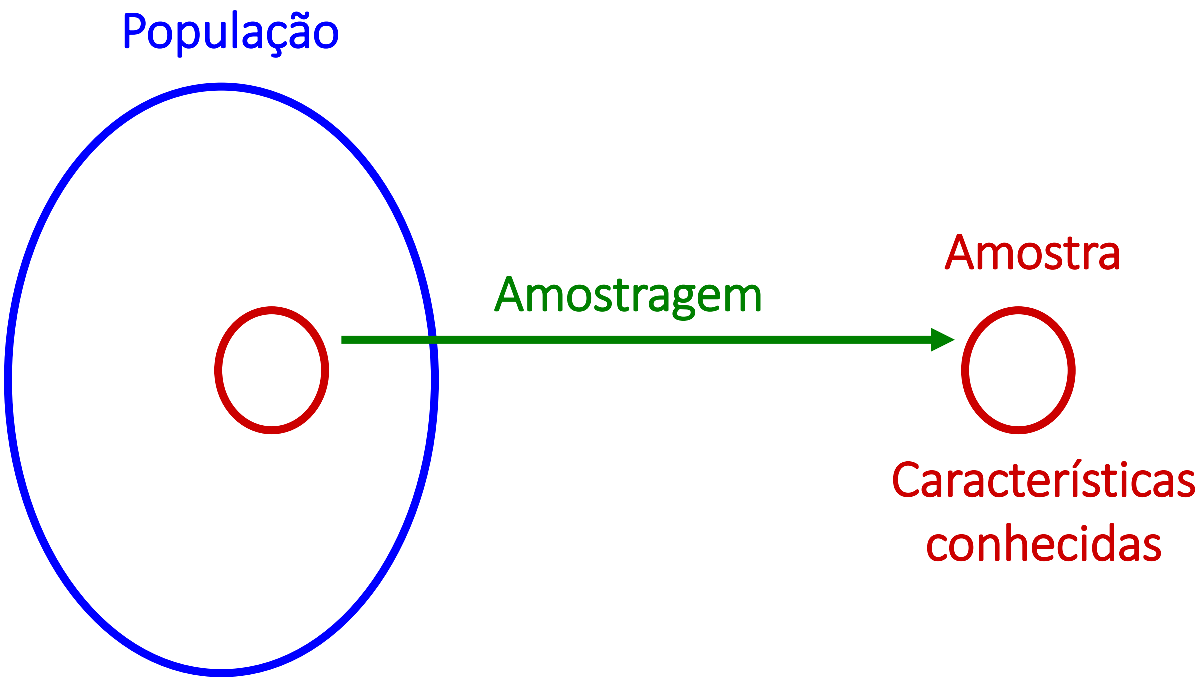
Você foi contratado para avaliar essa satisfação. Como você faria isso?

Qual o objetivo da pesquisa?

Determinar o nível médio de satisfação, da população dos clientes da rede de lojas

$\boldsymbol{\mu}$: *nível médio de satisfação da população de clientes da rede de lojas.*

Satisfação



Características desconhecidas

$$\boldsymbol{\mu} = ?$$

Amostra com $\boldsymbol{n = 5}$

1º cliente da amostra – Satisfação = 74

2º cliente da amostra – Satisfação = 72

3º cliente da amostra – Satisfação = 74

4º cliente da amostra – Satisfação = 68

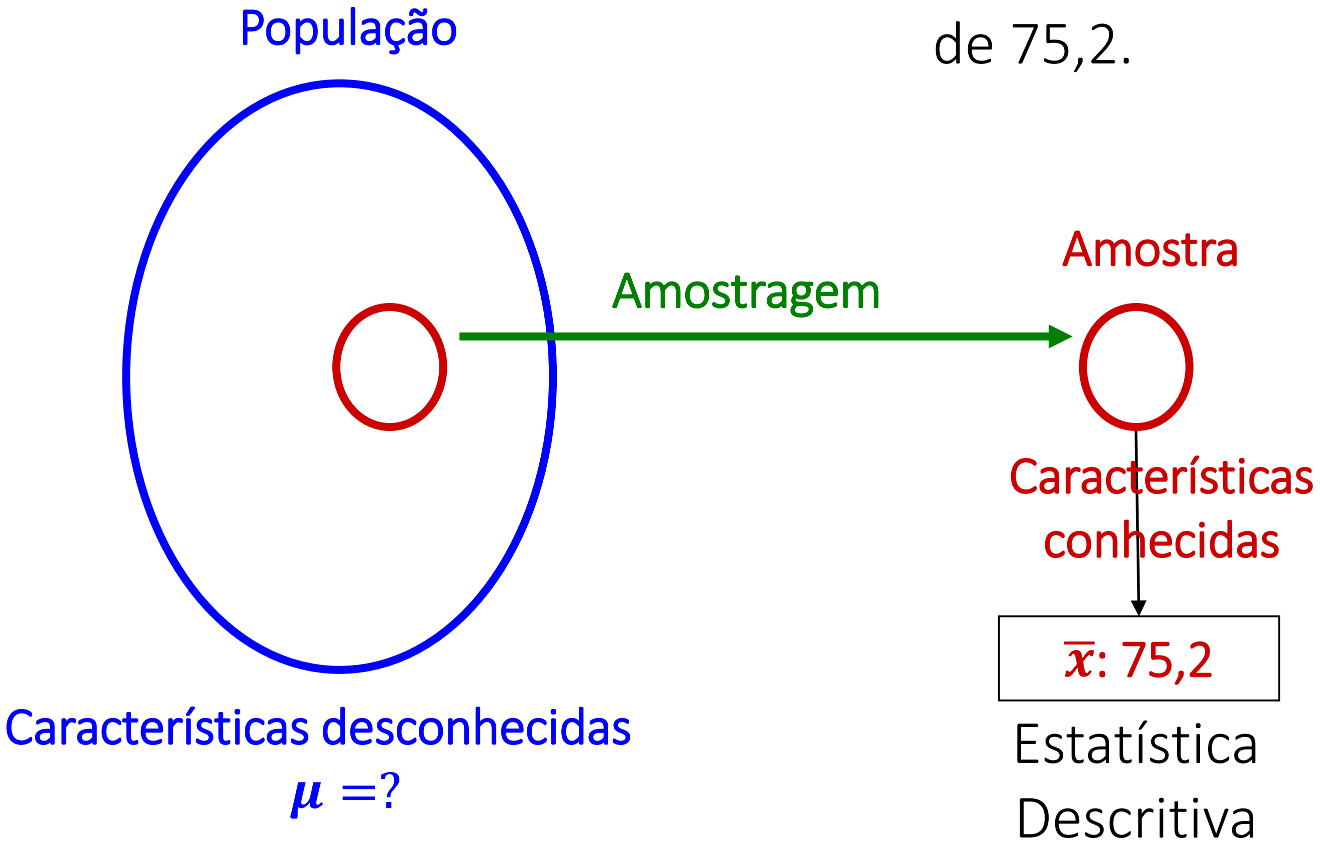
5º cliente da amostra – Satisfação = 88

Como obtenho informações sobre $\boldsymbol{\mu}$ a partir da amostra?

$$\bar{x} = \frac{74 + 72 + 74 + 68 + 88}{5} = 75{,}2$$

Satisfação

Foi extraída uma amostra de 5 clientes, obtendo-se uma média de 75,2.



Amostra com $n = 5$

Perguntas:

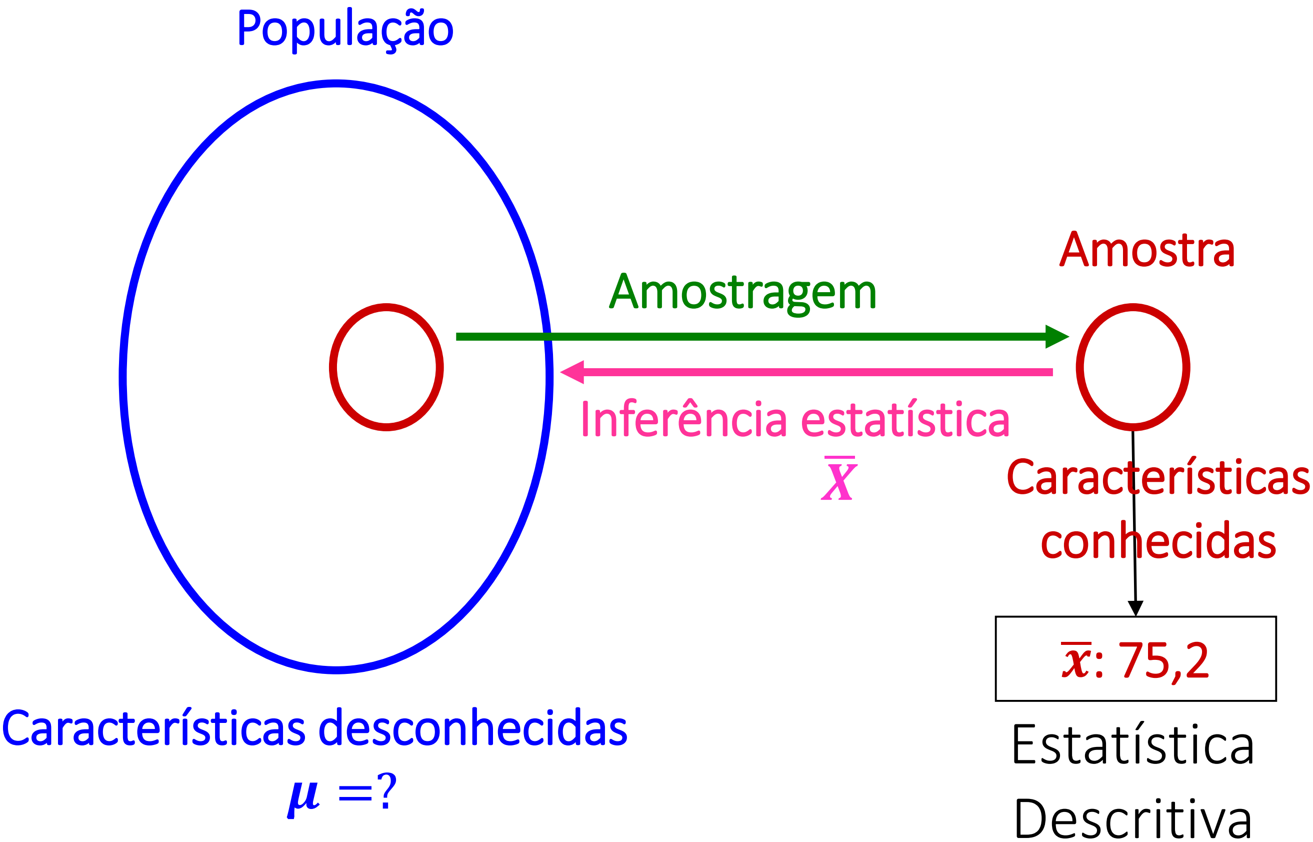
- $\bar{x}$ gera conhecimento sobre μ?
- μ vale exatamente o valor de $\bar{x}$?
- Se outra pessoa selecionar 5 clientes dessa população, então $\bar{x}$ também vale 75,2?
- Nesse caso, a partir de uma amostra de 5 clientes dessa população, a média amostral pode ser considerada uma variável aleatória?

Como obtenho informações sobre μ a partir da amostra?

$$\bar{x} = \frac{74 + 72 + 74 + 68 + 88}{5} = 75{,}2$$

Satisfação

Qual a relação entre $\bar{x}$ e μ?



Parâmetro e Estimador

Parâmetro: quantidade populacional, em geral, desconhecida.

No caso,

μ: satisfação média da população de clientes

Estimador: conta (função da amostra) que se faz para estimar o valor do parâmetro. No caso:

$$\bar{X} = \frac{\sum X_i}{n}$$

Estimativa: valor do estimador para uma particular amostra.

No caso:

$$\bar{x} = 75{,}2$$

1 amostra com $n = 5$

	A	B
1	**Amostra**	**1**
2	**1o cliente**	74
3	**2o cliente**	72
4	**3o cliente**	74
5	**4o cliente**	68
6	**5o cliente**	88
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		

1000 amostras com $n = 5$

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**Amostra**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
2	**1o cliente**	74	60	73	71	70	71	67	74	76	79	70	80	80	75	71
3	**2o cliente**	72	82	54	82	59	79	67	69	76	67	66	78	76	66	62
4	**3o cliente**	74	92	61	84	54	81	66	75	70	60	80	78	56	78	64
5	**4o cliente**	68	66	61	84	61	77	60	71	65	78	74	73	77	69	58
6	**5o cliente**	88	79	82	75	68	70	58	81	74	64	68	63	69	76	73
7																
8																
9																
10																
11																
12																
13																
14																
15																
16																

1000 amostras com $n = 5$

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**Amostra**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
2	**1o cliente**	74	60	73	71	70	71	67	74	76	79	70	80	80	75	71
3	**2o cliente**	72	82	54	82	59	79	67	69	76	67	66	78	76	66	62
4	**3o cliente**	74	92	61	84	54	81	66	75	70	60	80	78	56	78	64
5	**4o cliente**	68	66	61	84	61	77	60	71	65	78	74	73	77	69	58
6	**5o cliente**	88	79	82	75	68	70	58	81	74	64	68	63	69	76	73
7																
8	**média amostral**	75,2	75,8	66,2	79,2	62,4	75,6	63,6	74,0	72,2	69,6	71,6	74,4	71,6	72,8	65,
9																
10																
11																
12																
13																
14																
15																
16																
17																
18																

Estatísticas e distribuições amostrais

Uma **estatística** é qualquer função baseada nas observações de uma amostra aleatória.

Usamos estatísticas quando, para fins práticos, não é possível ter acesso as informações de toda uma população.

A distribuição de probabilidades de uma estatística é chamada de **distribuição amostral**.

TLC_Interativo

- Download do notebook
- Fazer individual e discutir em grupo
- Usar arquivo:

`Aula18_TLC_interativo.ipynb`

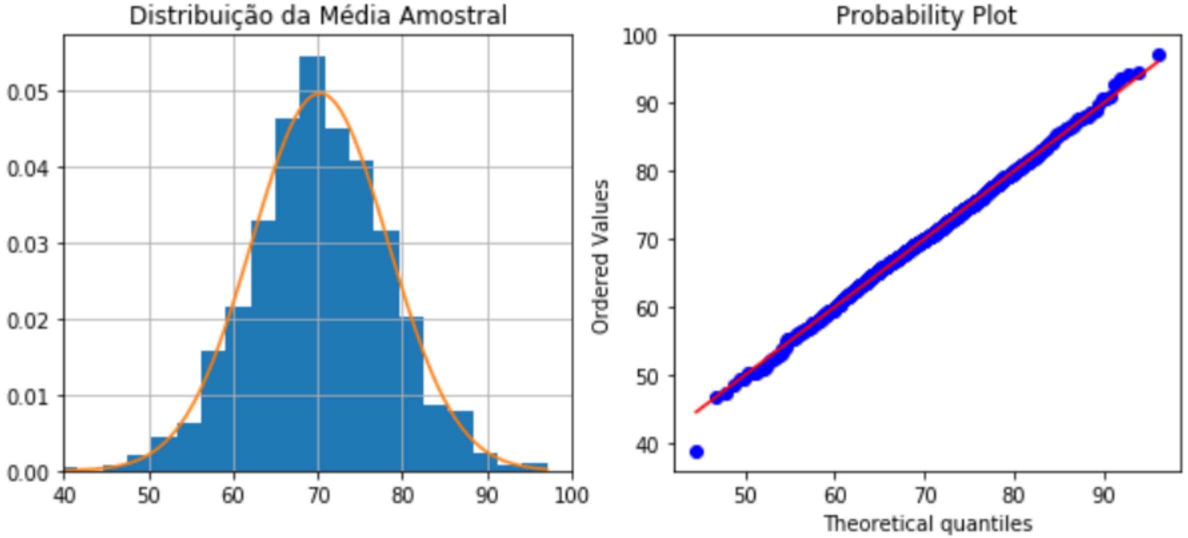
O que TLC_Interativo está fazendo?

Para mexer nesse arquivo, entenda que:

- A escolha da **distribuição** representa a distribuição da sua variável de interesse X.
- O valor de **n** representa o tamanho da sua amostra.

O que TLC_Interativo está fazendo?

Para mexer nesse arquivo, entenda que os gráficos construídos apresentam a **distribuição das médias amostrais** calculadas com amostra do valor de **n** escolhido.



O que TLC_Interativo está fazendo?

Por fim, os valores abaixo representam:

- Primeira linha mostra média de X e de $\bar{X}$
- Segunda linha mostra variância de X e de $\bar{X}$

```
('CONCLUSÃO DA SIMULAÇÃO:',
 '    Quando a média populacional for E(X)=70.500, então a média das médias amostrais será E(Xbarra)=70.529 ',
 '    Quando a variância populacional for Var(X)=64.000, então a variância das médias amostrais será Var(Xbarra)=67.488')
```

O que TLC_Interativo está fazendo?

Alterando as possíveis distribuições para X e também mexendo no valor de n, qual a sua conclusão em termos de:

- Distribuição
- Média e
- Variância

para média amostral $\overline{X}$.

Faça as suas conclusões.

Processo Interativo – Quando $X \sim Normal$ ➔ $\bar{X} \sim Normal$

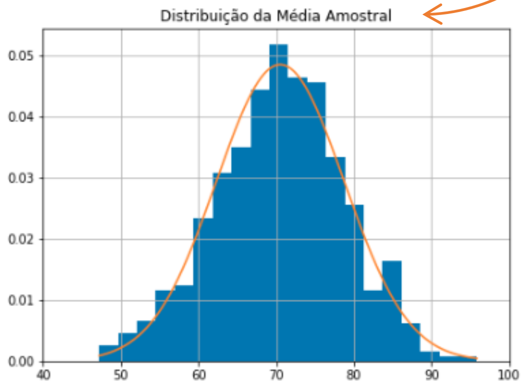

```
INPUT DA SIMULAÇÃO:
   Escolha a distribuição da variável de interesse X e um valor para o tamanho da amostra:
```

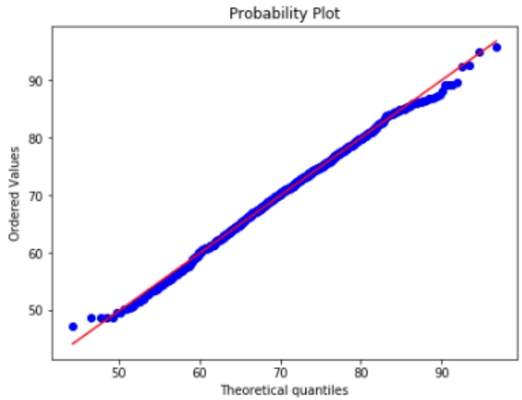
distribuição norm

n 1

$\bar{X}$ é normal quando n é pequeno

```
OUTPUT DA SIMULAÇÃO:
   Quando X ~ norm com tamanho da amostra igual a 1, analise a distribuição da média amostral:
```






```
('CONCLUSÃO DA SIMULAÇÃO:',
 '   Quando a média populacional for E(X)=70.500, então a média das médias amostrais será E(Xbarra)=70.529 ',
 '   Quando a variância populacional for Var(X)=64.000, então a variância das médias amostrais será Var(Xbarra)=67.488')
```

Processo Interativo – Quando $X \sim Normal$ ➔ $\bar{X} \sim Normal$

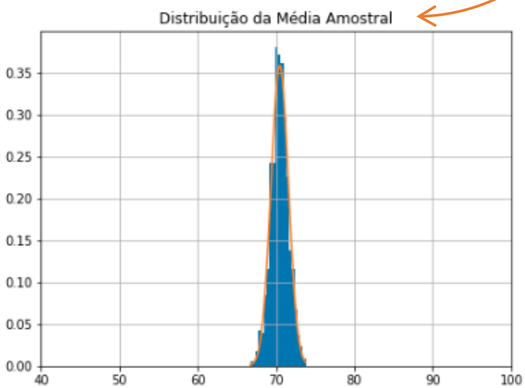

```
INPUT DA SIMULAÇÃO:
   Escolha a distribuição da variável de interesse X e um valor para o tamanho da amostra:
```

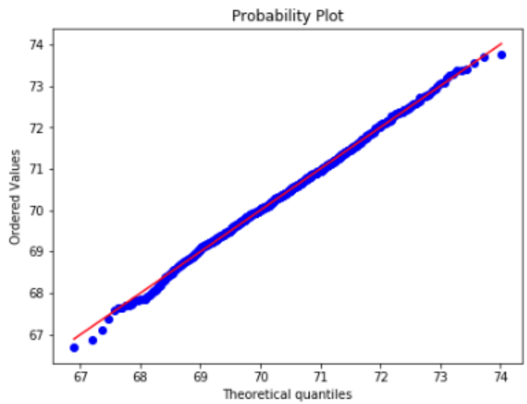
distribuição: norm

n: 50

$\bar{X}$ é normal quando n é grande

```
OUTPUT DA SIMULAÇÃO:
   Quando X ~ norm com tamanho da amostra igual a 50, analise a distribuição da média amostral:
```






```
('CONCLUSÃO DA SIMULAÇÃO:',
 '    Quando a média populacional for E(X)=70.500, então a média das médias amostrais será E(Xbarra)=70.456 ',
 '    Quando a variância populacional for Var(X)=64.000, então a variância das médias amostrais será Var(Xbarra)=1.240')
```

Distribuição amostral da média

$$X \sim Normal \xRightarrow{exata} \bar{X} \sim Normal$$

Assuma que X seja uma variável aleatória com média μ e variância σ^2, ou seja, $E(X) = \mu$ e $Var(X) = \sigma^2$.

Dado X_i uma variável aleatória *i.i.d.* (independente e identicamente distribuída) a X com distribuição **normal**, tem-se que:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

para qualquer n, essa distribuição é exata normal.

Processo Interativo – Quando $X \sim Uniforme \rightarrow \bar{X} \sim ?$

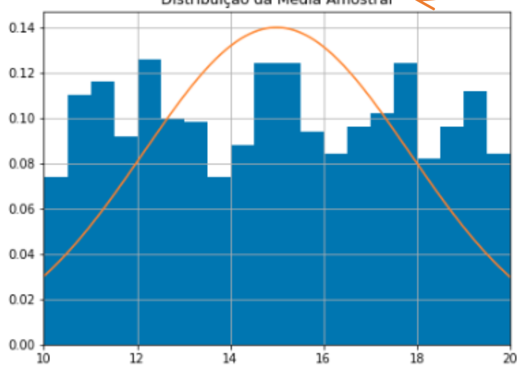

```
INPUT DA SIMULAÇÃO:
   Escolha a distribuição da variável de interesse X e um valor para o tamanho da amostra:
```

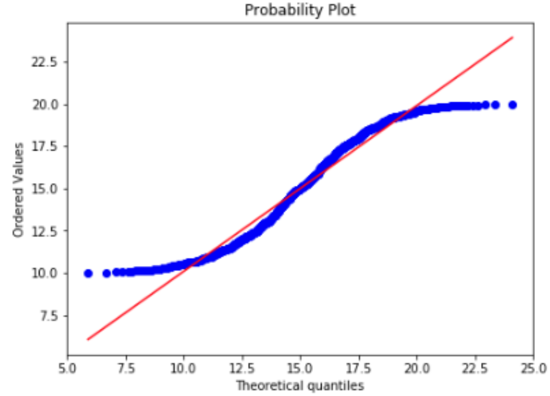
distribuição uniform

n 1

$\bar{X}$ não se aproxima da normal quando n é pequeno

```
OUTPUT DA SIMULAÇÃO:
   Quando X ~ uniform com tamanho da amostra igual a 1, analise a distribuição da média amostral:
```






```
('CONCLUSÃO DA SIMULAÇÃO:',
 '   Quando a média populacional for E(X)=15.000, então a média das médias amostrais será E(Xbarra)=14.990 ',
 '   Quando a variância populacional for Var(X)=8.333, então a variância das médias amostrais será Var(Xbarra)=8.105')
```

Processo Interativo –
Quando $X \sim Uniforme$ ➔ $\bar{X} \sim ?$

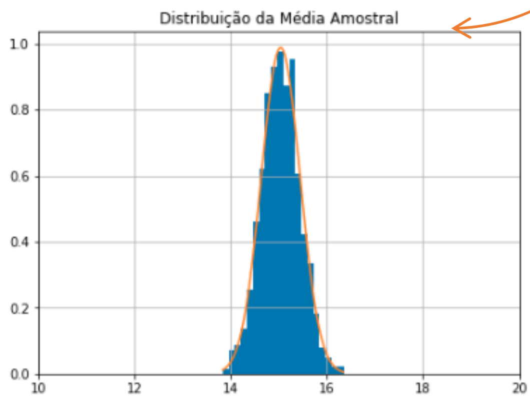
INPUT DA SIMULAÇÃO:
Escolha a distribuição da variável de interesse X e um valor para o tamanho da amostra:

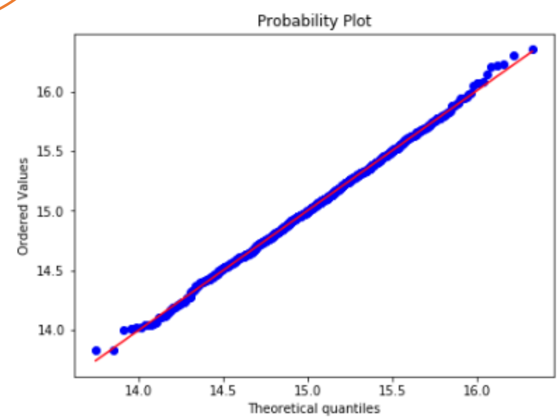
distribuição uniform

n 50

$\bar{X}$ se aproximada normal quando n é grande

OUTPUT DA SIMULAÇÃO:
Quando X ~ uniform com tamanho da amostra igual a 50, análise a distribuição da média amostral:





('CONCLUSÃO DA SIMULAÇÃO:',
' Quando a média populacional for E(X)=15.000, então a média das médias amostrais será E(Xbarra)=15.039 ',
' Quando a variância populacional for Var(X)=8.333, então a variância das médias amostrais será Var(Xbarra)=0.163')

Processo Interativo –
Quando $X \sim Exponencial$ ➔ $\bar{X} \sim ?$

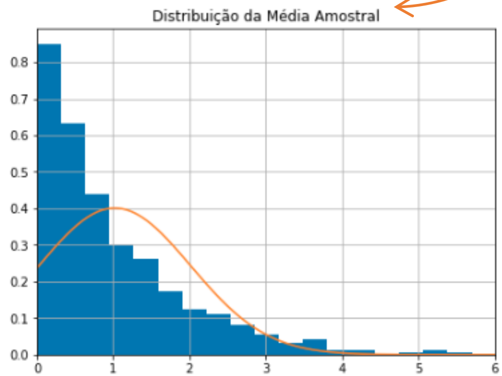

```
INPUT DA SIMULAÇÃO:
   Escolha a distribuição da variável de interesse X e um valor para o tamanho da amostra:
```

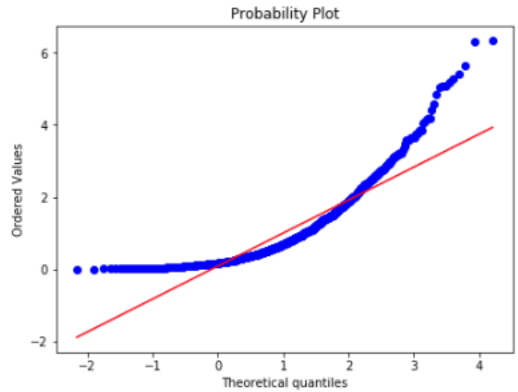
distribuição: expon

n: 1

$\bar{X}$ não se aproxima da normal quando n é pequeno

```
OUTPUT DA SIMULAÇÃO:
   Quando X ~ expon com tamanho da amostra igual a 1, analise a distribuição da média amostral:
```






```
('CONCLUSÃO DA SIMULAÇÃO:',
 '   Quando a média populacional for E(X)=1.000, então a média das médias amostrais será E(Xbarra)=1.019 ',
 '   Quando a variância populacional for Var(X)=1.000, então a variância das médias amostrais será Var(Xbarra)=0.990')
```

Processo Interativo –
Quando $X \sim Exponencial$ ➔ $\bar{X} \sim ?$

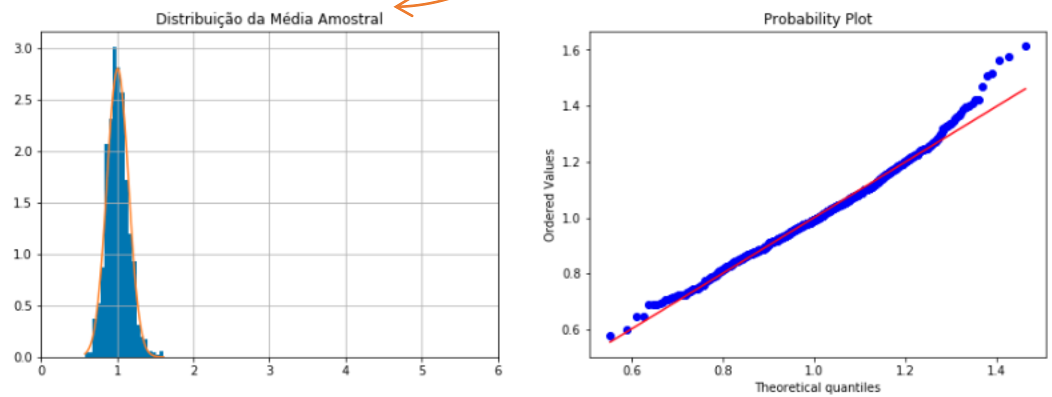

```
INPUT DA SIMULAÇÃO:
   Escolha a distribuição da variável de interesse X e um valor para o tamanho da amostra:
```

distribuição: expon

n: 50

$\bar{X}$ se aproximada normal quando n é grande

```
OUTPUT DA SIMULAÇÃO:
   Quando X ~ expon com tamanho da amostra igual a 50, analise a distribuição da média amostral:
```




```
('CONCLUSÃO DA SIMULAÇÃO:',
 '   Quando a média populacional for E(X)=1.000, então a média das médias amostrais será E(Xbarra)=1.008 ',
 '   Quando a variância populacional for Var(X)=1.000, então a variância das médias amostrais será Var(Xbarra)=0.020')
```

Distribuição amostral da média

$$X \sim ? \overset{TLC}{\Longrightarrow} \bar{X} \sim Normal$$

Teorema do Limite Central:

Assuma que X seja uma variável aleatória com média μ e variância σ^2, ou seja, $E(X) = \mu$ e $Var(X) = \sigma^2$.

Dado X_i uma variável aleatória *i.i.d.* (independente e identicamente distribuída) a X com distribuição **qualquer**, tem-se que:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

quando n for suficientemente grande.

Obs: Em alguns casos específicos, pode-se considerar X_i com distribuições diferentes ou com certa dependência entre elas.