



# ME951 - Estatística e Probabilidade I

Parte 14

1S2024

# Distribuição Amostral e Teorema Central do Limite

# Estimar uma proporção: Eleições para a prefeitura

- Quero saber se o candidato  $A$  vai ganhar as eleições para prefeito.
- Quero saber o parâmetro populacional  $p =$  proporção de pessoas que votam em  $A$ .
- Posso esperar o resultado das eleições para saber, ou seja, teríamos as respostas de todas as pessoas da cidade.
- Posso usar uma amostra para estimar a proporção de votos para  $A$ .
- Quão boa é a estimativa? É precisa?
- Posso pensar no problema de duas formas: Modo 1 e Modo 2.



# Modo 1

- Cidade com  $N$  pessoas.
- Considere

$X_i = 1$ , se a pessoa  $i$  vota em  $A$

$X_i = 0$ , se a pessoa  $i$  não vota em  $A$ .

- $\mathbf{X} = (X_1, X_2, \dots, X_N)$ : respostas de toda a população (temos no dia da eleição).
- Média populacional:

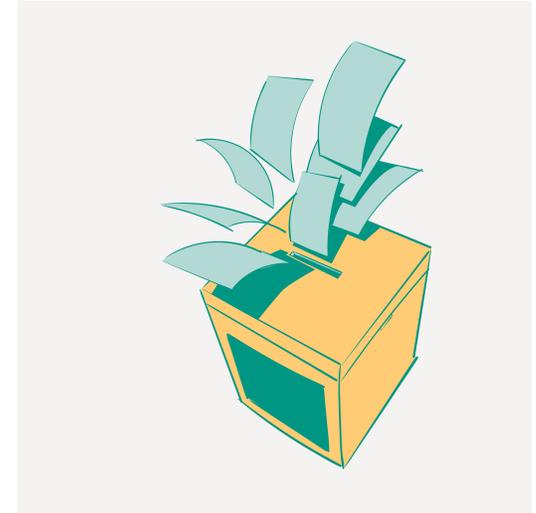
$$p = \frac{1}{N} \sum_{i=1}^N X_i$$



# Modo 1

- Variância populacional:

$$\begin{aligned}\sigma^2 &= \frac{1}{N} \sum_{i=1}^N (X_i - p)^2 \\ &= \frac{1}{N} \sum_{i=1}^N (X_i^2 - 2pX_i + p^2) \\ &= \frac{\sum_{i=1}^N X_i^2 - 2p \sum_{i=1}^N X_i + \sum_{i=1}^N p^2}{N} \\ &= \frac{\sum_{i=1}^N X_i - 2p \sum_{i=1}^N X_i + \sum_{i=1}^N p^2}{N} \\ &= \frac{Np - 2pNp + Np^2}{N} = p(1 - p)\end{aligned}$$



# Modo 1

Vimos que,

- $p$  = proporção de pessoas que votam em  $A$  na cidade;
- $\sigma^2 = p(1 - p)$  é a variância da população.

Até o dia da eleição, não sabemos  $p$ .

Coletamos uma amostra aleatória de tamanho  $n$  para uma pesquisa eleitoral.

$\hat{p}$  : proporção de pessoas que votam em  $A$  na amostra.

- Quão boa é a estimativa? É precisa?
- Se outra pessoa também coleta uma amostra aleatória de tamanho  $n$  e calcula  $\hat{p}$  teremos o mesmo valor?

# Modo 1 - Exemplo $N = 5$

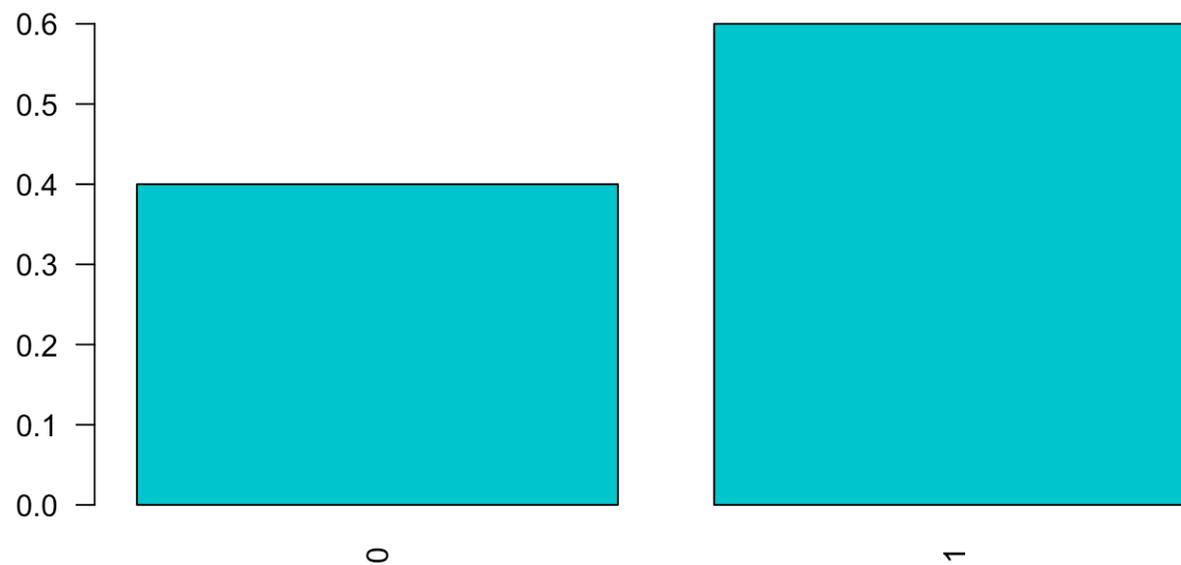
•  $\mathbf{X} = (X_1, X_2, \dots, X_5) = (1, 0, 1, 0, 1)$

$$p = \frac{\sum_{i=1}^5 X_i}{5} = \frac{3}{5} = 0.6$$

$$\begin{aligned}\sigma^2 &= \frac{1}{5} \sum_{i=1}^N (X_i - p)^2 \\ &= \frac{3 \times (1 - 0.6)^2 + 2 \times (0 - 0.6)^2}{5} \\ &= 0.24 \\ &= p(1 - p)\end{aligned}$$

# Modo 1 - Exemplo $N = 5$

Gráfico de barras (proporção) dos dados populacionais:



# Modo 1 - Exemplo $N = 5$ e $n = 2$

- $N^n = 25$  amostras possíveis

$$\mathbf{X} = (X_1, X_2, \dots, X_5) = (1, 0, 1, 0, 1)$$

	Pessoa amostrada 1	Pessoa amostrada 2	$\hat{p}$		Pessoa amostrada 1	Pessoa amostrada 2	$\hat{p}$
1	1	1	1.0	14	4	3	0.5
2	2	1	0.5	15	5	3	1
3	3	1	1.0	16	1	4	0.5
4	4	1	0.5	17	2	4	0
5	5	1	1.0	18	3	4	0.5
6	1	2	0.5	19	4	4	0
7	2	2	0.0	20	5	4	0.5
8	3	2	0.5	21	1	5	1
9	4	2	0.0	22	2	5	0.5
10	5	2	0.5	23	3	5	1
11	1	3	1.0	24	4	5	0.5
12	2	3	0.5	25	5	5	1
13	3	3	1.0				

# Modo 1 - Exemplo $N = 5$ e $n = 2$

Distribuição amostral de  $\hat{p}$ :

$x$	$P(\hat{p} = x)$
0	0.16
0.5	0.48
1	0.36

$$E(\hat{p}) = 0 \times 0.16 + 0.5 \times 0.48 + 1 \times 0.36 = 0.6 = p$$

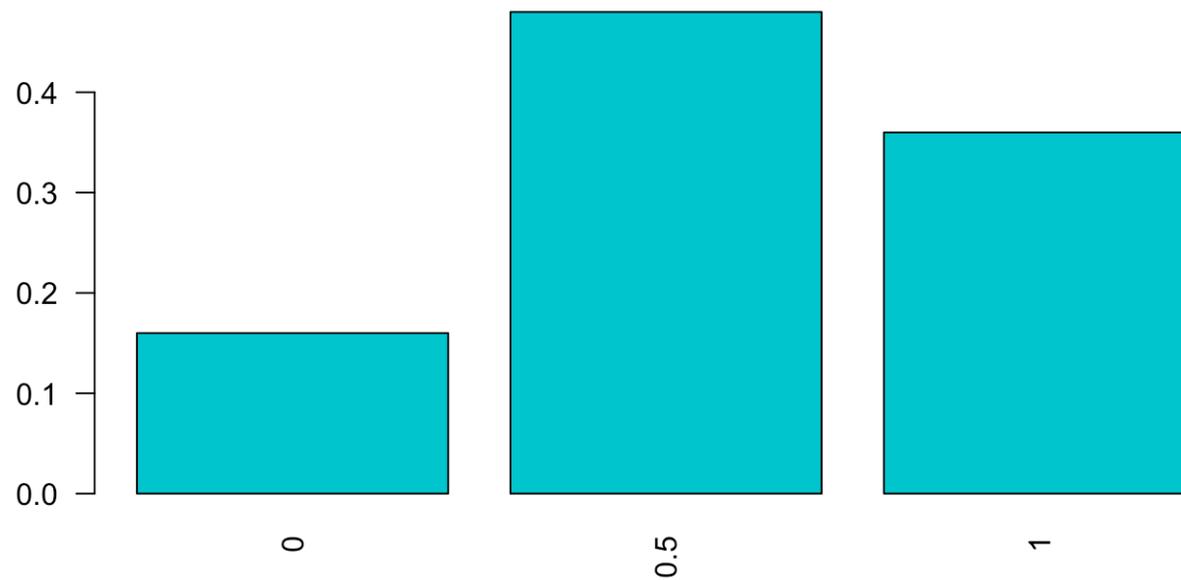
$$Var(\hat{p}) = E[(\hat{p} - p)^2]$$

$$= 0.16 \times (0 - 0.6)^2 + 0.48 \times (0.5 - 0.6)^2 + 0.36 \times (1 - 0.6)^2$$

$$= 0.12 = \frac{0.24}{2} = \frac{p(1-p)}{n}$$

# Modo 1 - Exemplo $N = 5$ e $n = 2$

Distribuição amostral de  $\hat{p}$ :



# Modo 1 - Exemplo $N = 5$ e $n = 3$

$N^n = 125$  amostras possíveis.

	Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\hat{p}$		Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\hat{p}$
1	1	1	1	1.000	14	4	3	1	0.667
2	2	1	1	0.667	15	5	3	1	1.000
3	3	1	1	1.000	16	1	4	1	0.667
4	4	1	1	0.667	17	2	4	1	0.333
5	5	1	1	1.000	18	3	4	1	0.667
6	1	2	1	0.667	19	4	4	1	0.333
7	2	2	1	0.333	20	5	4	1	0.667
8	3	2	1	0.667	21	1	5	1	1.000
9	4	2	1	0.333	22	2	5	1	0.667
10	5	2	1	0.667	23	3	5	1	1.000
11	1	3	1	1.000	24	4	5	1	0.667
12	2	3	1	0.667	25	5	5	1	1.000
13	3	3	1	1.000	26	1	1	2	0.667

# Modo 1 - Exemplo $N = 5$ e $n = 3$

	Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\hat{p}$		Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\hat{p}$
27	2	1	2	0.333	42	2	4	2	0.000
28	3	1	2	0.667	43	3	4	2	0.333
29	4	1	2	0.333	44	4	4	2	0.000
30	5	1	2	0.667	45	5	4	2	0.333
31	1	2	2	0.333	46	1	5	2	0.667
32	2	2	2	0.000	47	2	5	2	0.333
33	3	2	2	0.333	48	3	5	2	0.667
34	4	2	2	0.000	49	4	5	2	0.333
35	5	2	2	0.333	50	5	5	2	0.667
36	1	3	2	0.667	51	1	1	3	1.000
37	2	3	2	0.333	52	2	1	3	0.667
38	3	3	2	0.667	53	3	1	3	1.000
39	4	3	2	0.333	54	4	1	3	0.667
40	5	3	2	0.667	55	5	1	3	1.000
41	1	4	2	0.333	56	1	2	3	0.667

# Modo 1 - Exemplo $N = 5$ e $n = 3$

	Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\hat{p}$		Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\hat{p}$
57	2	2	3	0.333	72	2	5	3	0.667
58	3	2	3	0.667	73	3	5	3	1.000
59	4	2	3	0.333	74	4	5	3	0.667
60	5	2	3	0.667	75	5	5	3	1.000
61	1	3	3	1.000	76	1	1	4	0.667
62	2	3	3	0.667	77	2	1	4	0.333
63	3	3	3	1.000	78	3	1	4	0.667
64	4	3	3	0.667	79	4	1	4	0.333
65	5	3	3	1.000	80	5	1	4	0.667
66	1	4	3	0.667	81	1	2	4	0.333
67	2	4	3	0.333	82	2	2	4	0.000
68	3	4	3	0.667	83	3	2	4	0.333
69	4	4	3	0.333	84	4	2	4	0.000
70	5	4	3	0.667	85	5	2	4	0.333
71	1	5	3	1.000	86	1	3	4	0.667

# Modo 1 - Exemplo $N = 5$ e $n = 3$

	Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\hat{p}$		Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\hat{p}$
87	2	3	4	0.333	102	2	1	5	0.667
88	3	3	4	0.667	103	3	1	5	1.000
89	4	3	4	0.333	104	4	1	5	0.667
90	5	3	4	0.667	105	5	1	5	1.000
91	1	4	4	0.333	106	1	2	5	0.667
92	2	4	4	0.000	107	2	2	5	0.333
93	3	4	4	0.333	108	3	2	5	0.667
94	4	4	4	0.000	109	4	2	5	0.333
95	5	4	4	0.333	110	5	2	5	0.667
96	1	5	4	0.667	111	1	3	5	1.000
97	2	5	4	0.333	112	2	3	5	0.667
98	3	5	4	0.667	113	3	3	5	1.000
99	4	5	4	0.333	114	4	3	5	0.667
100	5	5	4	0.667	115	5	3	5	1.000
101	1	1	5	1.000	116	1	4	5	0.667

# Modo 1 - Exemplo $N = 5$ e $n = 3$

	Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\hat{p}$
117	2	4	5	0.333
118	3	4	5	0.667
119	4	4	5	0.333
120	5	4	5	0.667
121	1	5	5	1.000
122	2	5	5	0.667
123	3	5	5	1.000
124	4	5	5	0.667
125	5	5	5	1.000

# Modo 1 - Exemplo $N = 5$ e $n = 3$

Distribuição amostral de  $\hat{p}$ :

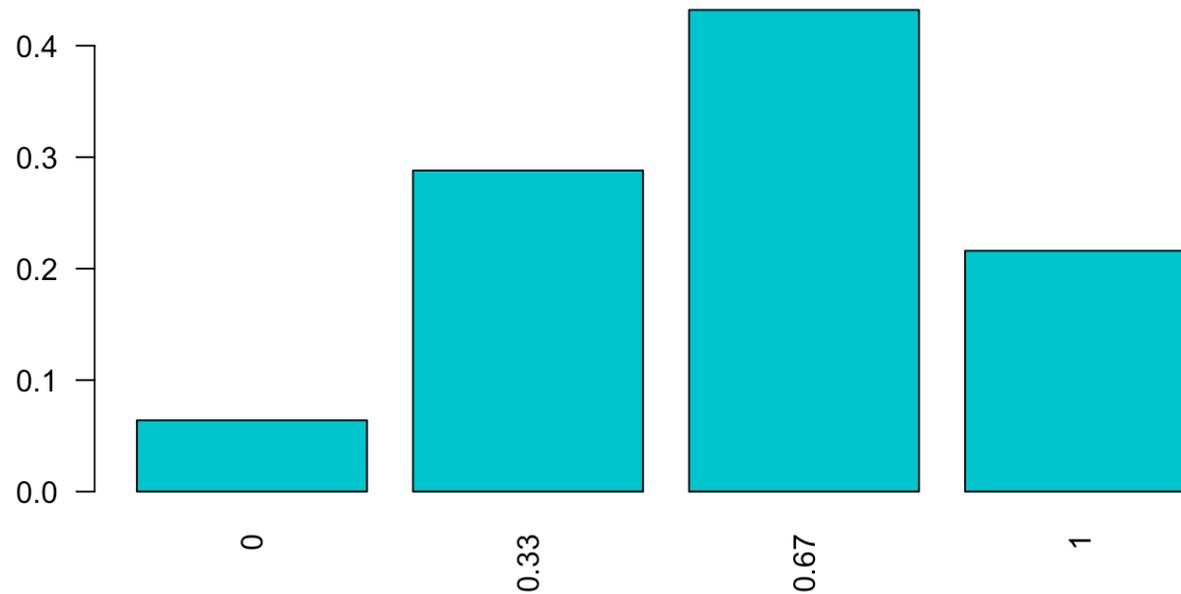
$x$	$P(\hat{p} = x)$
0	0.064
0.333	0.288
0.667	0.432
1	0.216

$$\begin{aligned} E(\hat{p}) &= 0 \times 0.064 + 0.333 \times 0.288 + 0.667 \times 0.432 + 1 \times 0.216 \\ &= 0.6 = p \end{aligned}$$

$$\text{Var}(\hat{p}) = E[(\hat{p} - p)^2] = 0.08 = \frac{0.24}{3} = \frac{p(1-p)}{n}$$

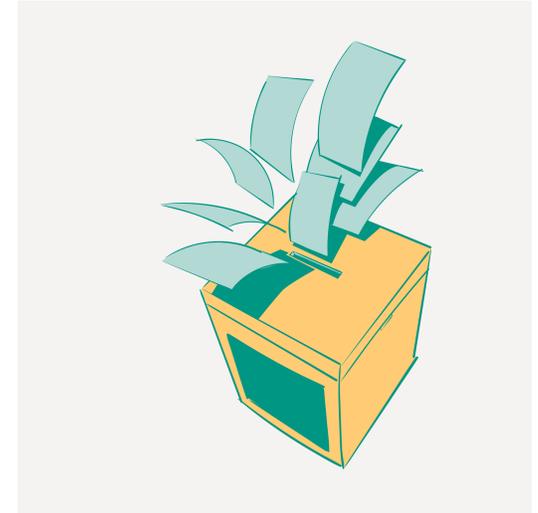
# Modo 1 - Exemplo $N = 5$ e $n = 3$

Distribuição amostral de  $\hat{p}$ :



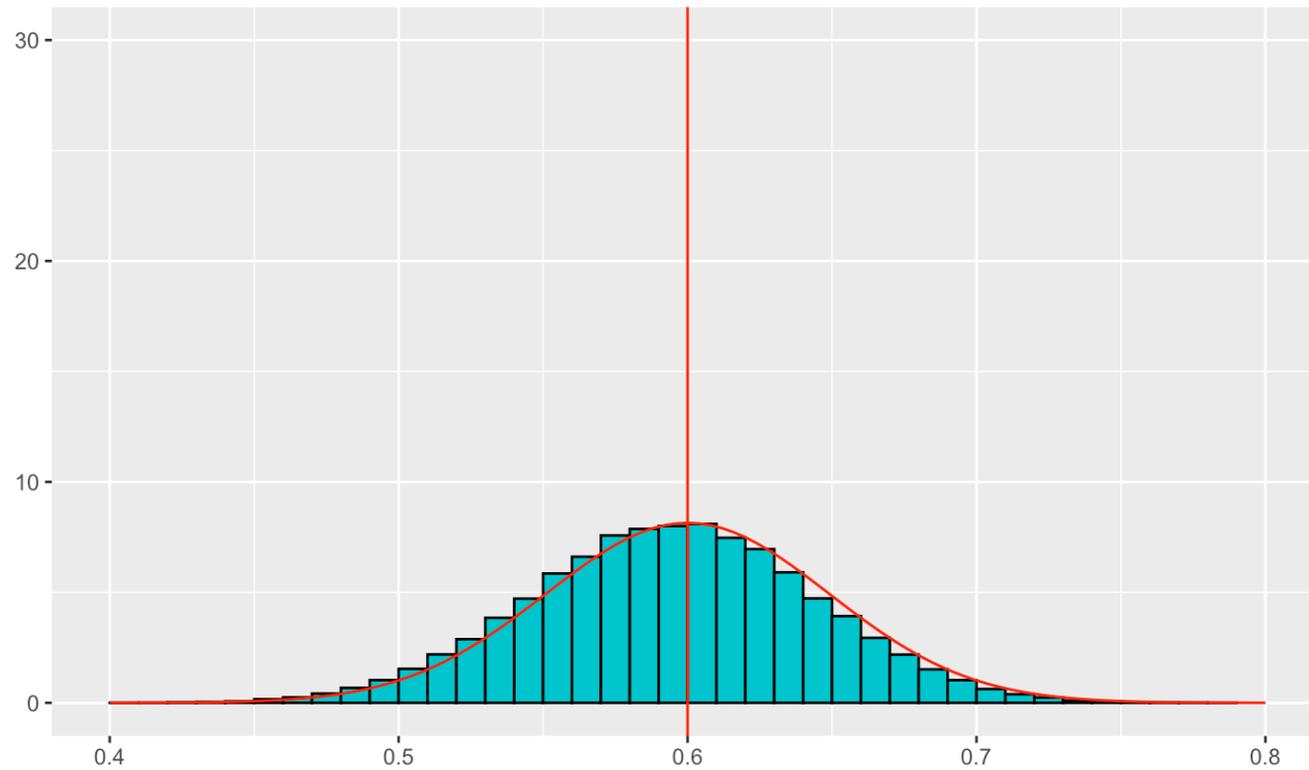
# Modo 1

- $\mathbf{X} = (X_1, \dots, X_N)$  é fixo
- Amostra aleatória de tamanho  $n$
- $\hat{p}$  é v.a. (pelo processo de amostragem)
- $E(\hat{p}) = p$
- $Var(\hat{p}) = \frac{p(1-p)}{n}$



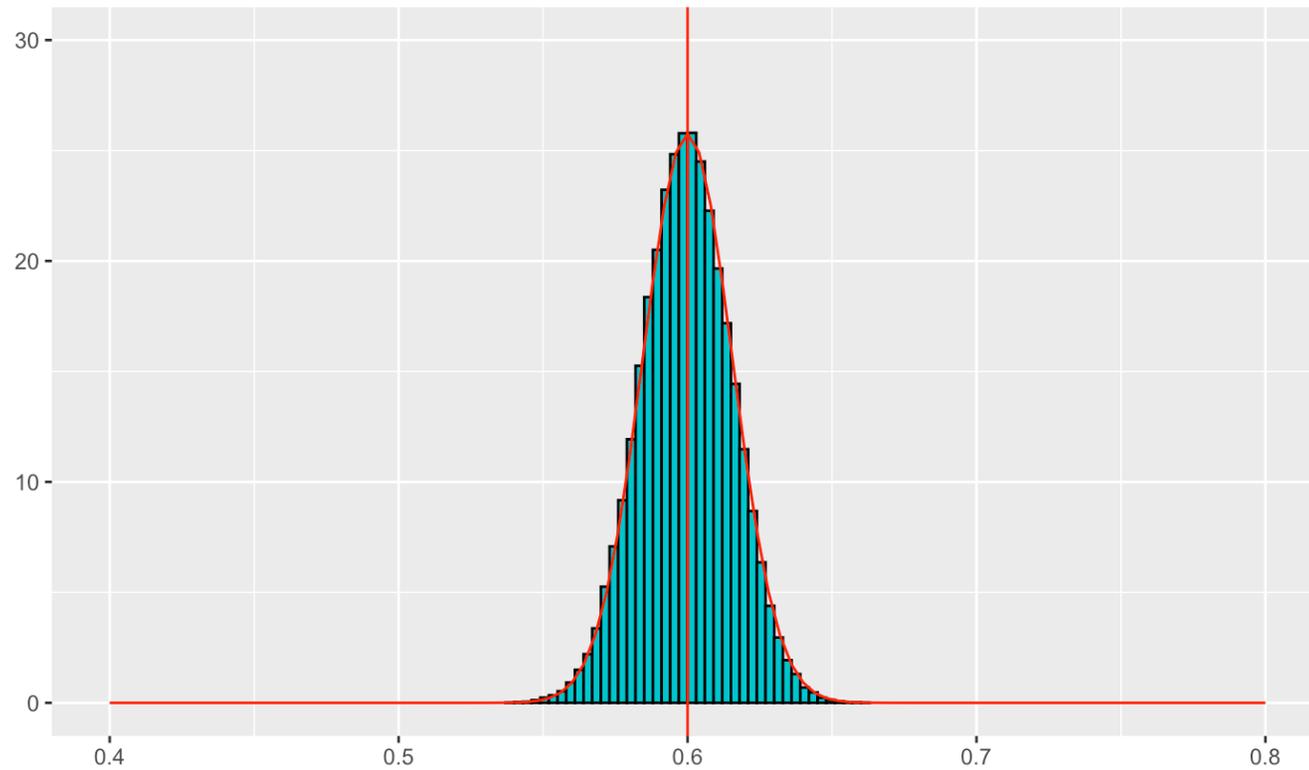
# Modo 1 - Exemplo $N = 1000000$ e $n = 100$

- $p = 0.6$ . Distribuição amostral de  $\hat{p}$ :



# Modo 1 - Exemplo $N = 1000000$ e $n = 1000$

- $p = 0.6$ . Distribuição amostral de  $\hat{p}$ :



## Modo 2

Suponha que a resposta de uma pessoa da cidade sobre se vota ou não no candidato  $A$  possa ser representada por uma **variável aleatória**. Seja  $X$  a v.a. que assume o valor 1 com probabilidade  $p$  ou 0 com probabilidade  $1 - p$ .

$$X \sim \text{Bernoulli}(p)$$

$$\begin{aligned}\mathbb{E}(X) &= 1 \times P(X = 1) + 0 \times P(X = 0) \\ &= 1 \times p + 0 \times (1 - p) = p\end{aligned}$$

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - p)^2] \\ &= (1 - p)^2 \times P(X = 1) + (0 - p)^2 \times P(X = 0) \\ &= p(1 - p)^2 + (1 - p)p^2 \\ &= p(1 - p)\end{aligned}$$

# Modo 2 - Exemplo $n = 2$

Todas as combinações possíveis de amostras com  $n = 2$  são:

Possibilidades	$(X_1 = 1, X_2 = 1)$	$(X_1 = 1, X_2 = 0)$	$(X_1 = 0, X_2 = 1)$	$(X_1 = 0, X_2 = 0)$
$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$	1	0.5	0.5	0
$P(X_1 = i, X_2 = j)$	$p^2$	$p(1-p)$	$(1-p)p$	$(1-p)^2$

$$\mathbb{E}(\hat{p}) = 1 \times p^2 + 0.5 \times p(1-p) + 0.5 \times (1-p)p + 0 \times (1-p)^2 = p$$

$$\text{Var}(\hat{p}) = \mathbb{E}[(\hat{p} - p)^2]$$

$$= (1-p)^2 \times p^2 + (0.5-p)^2 p(1-p) + (0.5-p)^2 (1-p)p + (0-p)^2 (1-p)^2$$

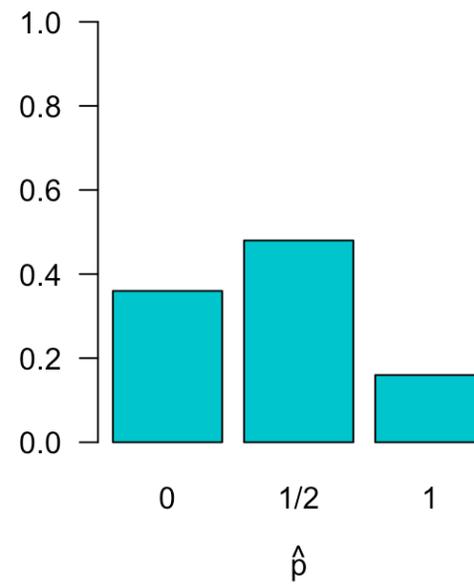
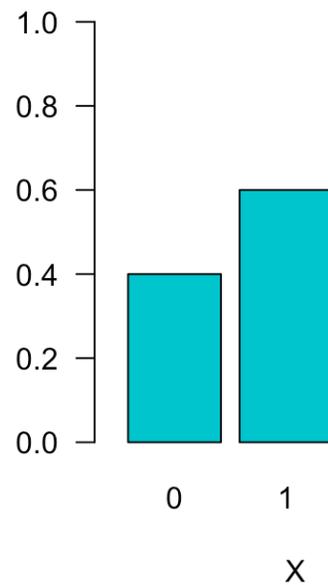
$$= \frac{p(1-p)}{2}$$

Note que:  $\mathbb{E}(\hat{p}) = p = \mathbb{E}(X)$  e  $\text{Var}(\hat{p}) = \frac{\text{Var}(X)}{n}$ .

# Modo 2 - Exemplo $n = 2$

•  $p = 0.6$

Gráficos das distribuições de probabilidade de  $X \sim \text{Bernoulli}(p = 0.6)$  e  $\hat{p}$ :



# Resultado

Seja  $X$  uma v.a. com distribuição de Bernoulli com parâmetro  $p$ . Sabe-se que  $E(X) = p$  e  $Var(X) = p(1 - p)$ . Considere uma amostra aleatória  $X_1, X_2, \dots, X_n$  de  $X$ .

A proporção amostral

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$$

tem as seguintes propriedades:

$$\mathbb{E}(\hat{p}) = p \quad \text{e} \quad Var(\hat{p}) = \frac{p(1 - p)}{n}.$$

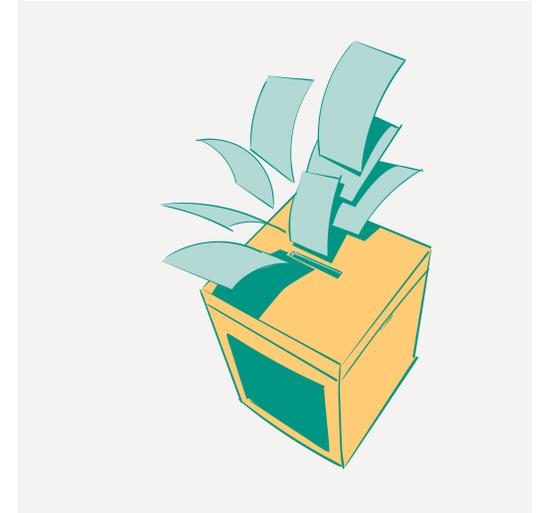
(propriedade de linearidade da esperança e da variância, esta última em caso de independência)

Ou seja, embora  $p$  seja desconhecido, sabemos que o valor esperado da proporção amostral é  $p$ .

Além disso, conforme o tamanho amostral aumenta, a imprecisão de  $\hat{p}$  para estimar  $p$  fica cada vez menor, pois  $Var(\hat{p}) = p(1 - p)/n$  é inversamente proporcional ao tamanho amostral  $n$ .

# Modo 2

- $X_i \sim \text{Bernoulli}(p)$  é v.a. (o voto ou não em  $A$  é considerado uma v.a.)
- Amostra aleatória de tamanho  $n$
- $\hat{p}$  é v.a. (é combinação linear de v.a.'s)
- $E(\hat{p}) = p$
- $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$



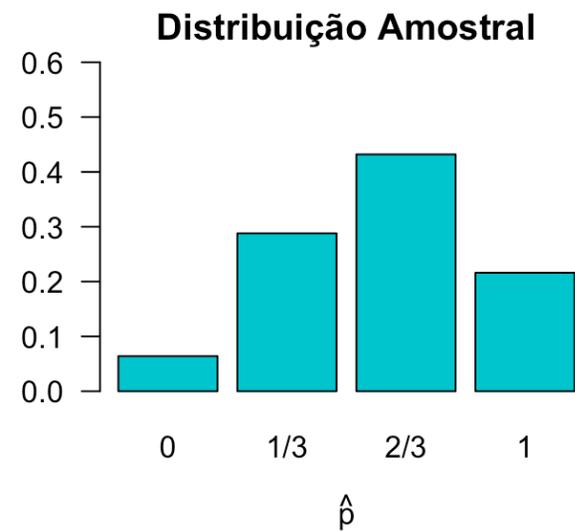
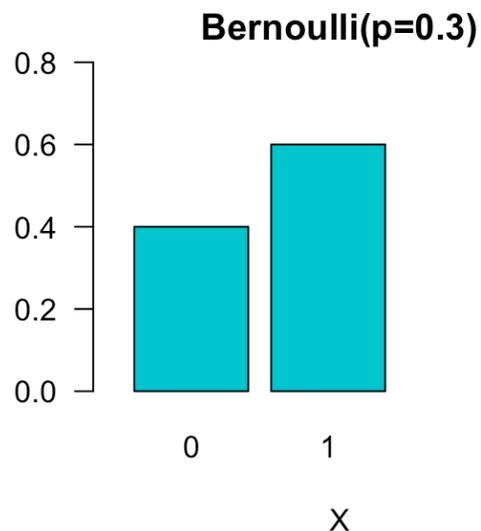
# Modo 2 - Exemplo $n = 3$

•  $p = 0.6$

Amostra aleatória  $n = 3$  de  $X \sim \text{Bernoulli}(p = 0.6)$ .

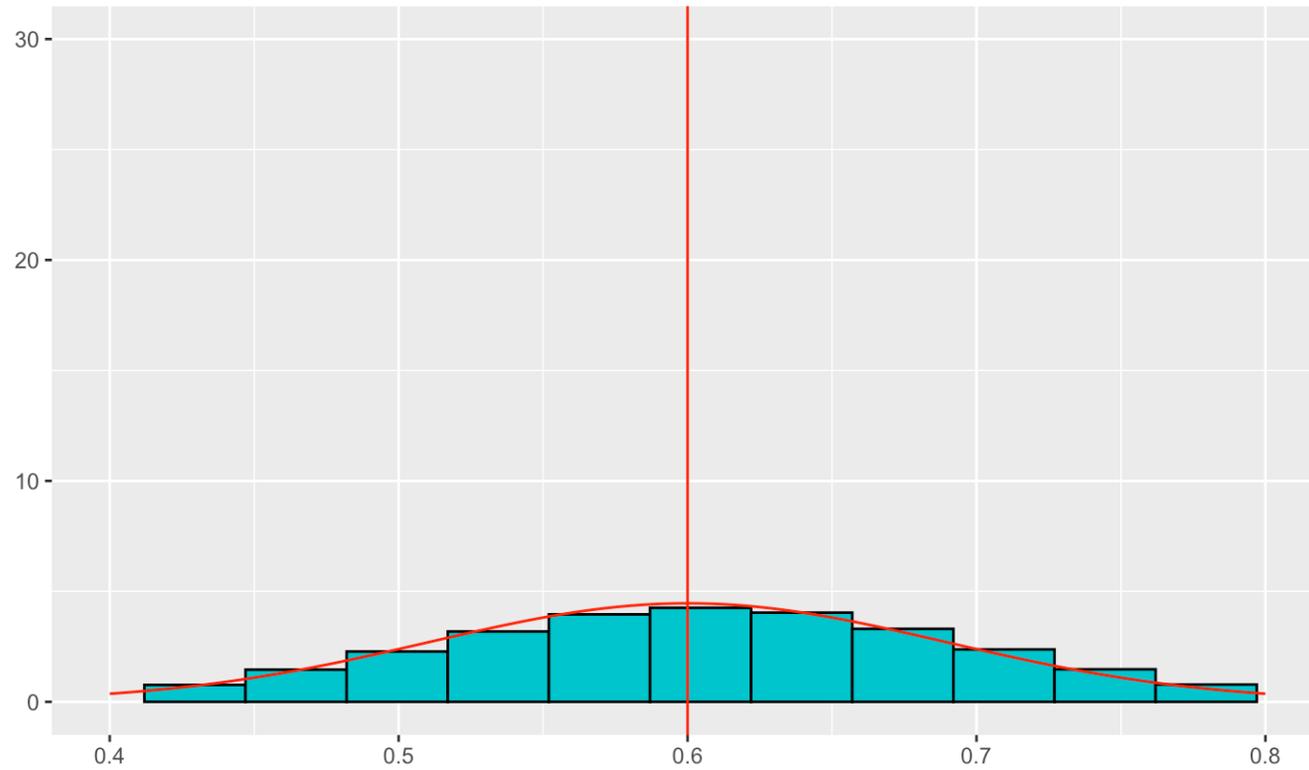
$$\mathbb{E}(X) = p = 0.6 \quad \Rightarrow \quad \mathbb{E}(\hat{p}) = 0.6$$

$$\text{Var}(X) = p(1 - p) = 0.24 \quad \Rightarrow \quad \text{Var}(\hat{p}) = \frac{0.24}{3} = 0.08$$



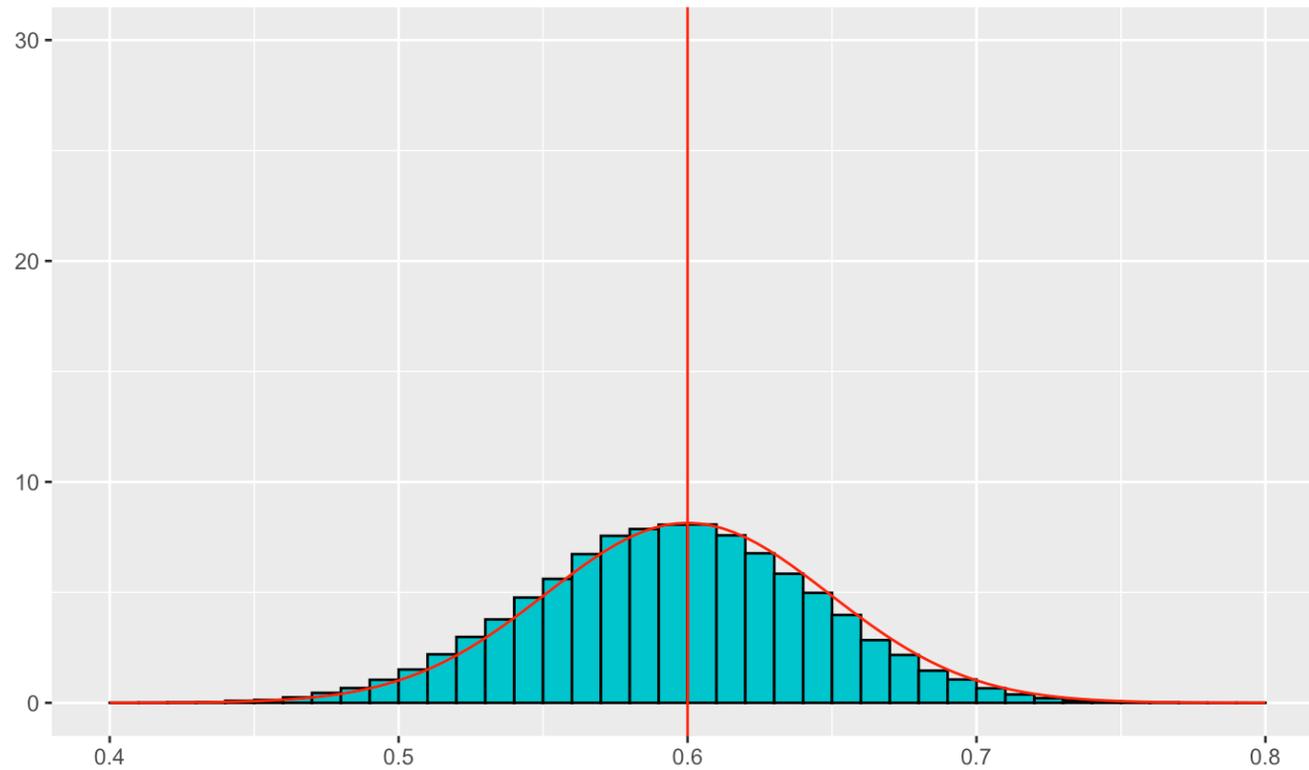
# Modo 2 - Exemplo $n = 30$

•  $p = 0.6$



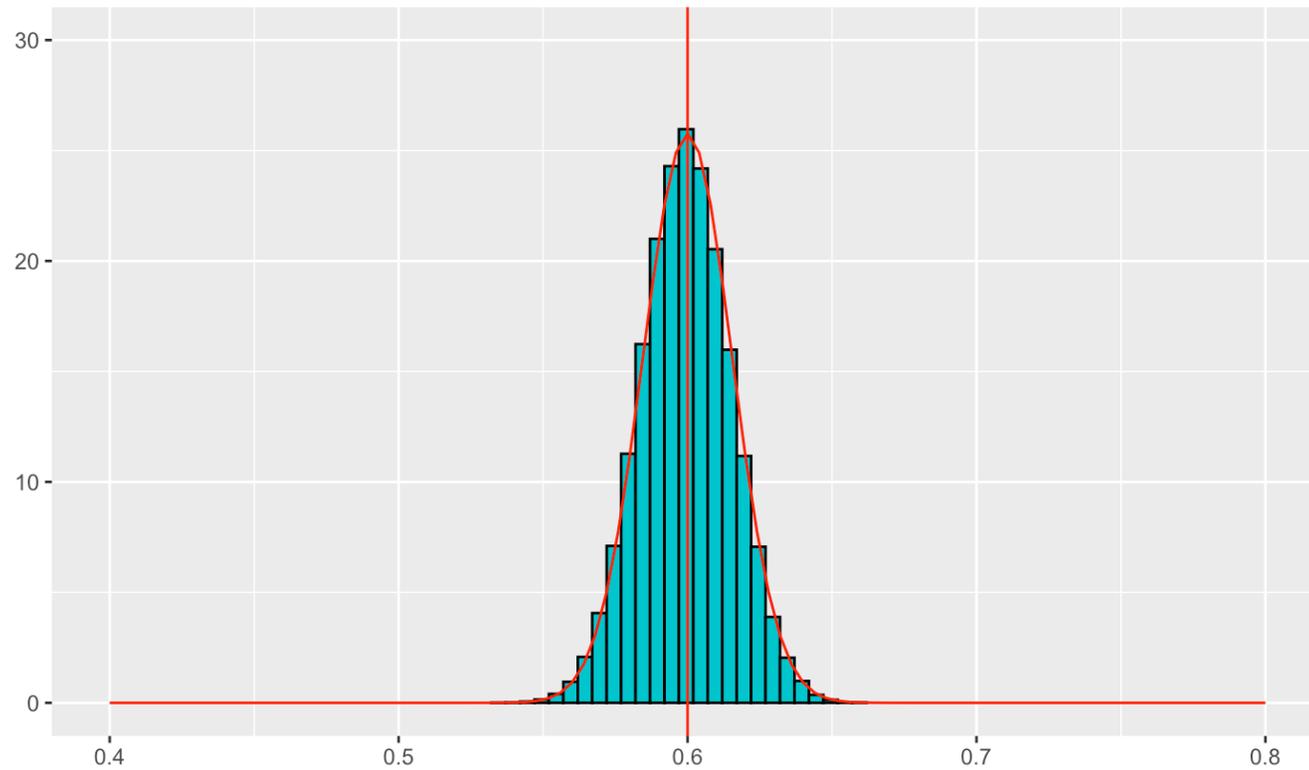
# Modo 2 - Exemplo $n = 100$

•  $p = 0.6$

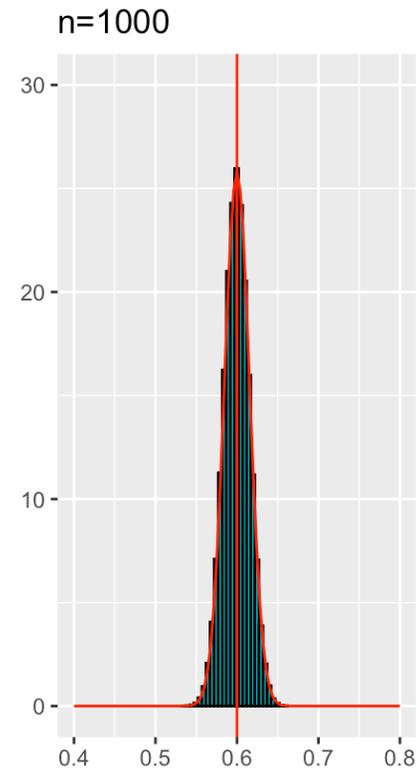
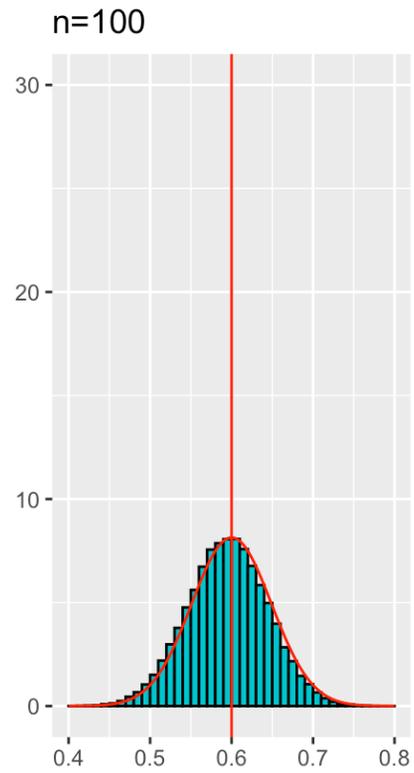
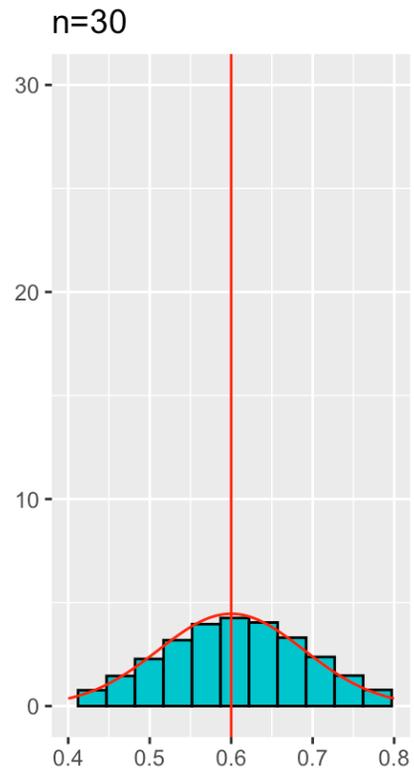


# Modo 2 - Exemplo $n = 1000$

•  $p = 0.6$



•  $p = 0.6$



# Resumo dos exemplos

- Modo 1: repostas são “fixas”, com média populacional  $p$  e variância populacional  $p(1 - p)$ .
- Modo 2: respostas são v.a.'s  $X \sim \text{Bernoulli}(p)$ ,  $E(X) = p$ ,  $\text{Var}(X) = p(1 - p)$ .
- Em ambos os casos, a partir do momento que retiro uma amostra aleatória de tamanho  $n$ , temos as mesmas propriedades e comportamento para a proporção amostral  $\hat{p}$ :  $E(\hat{p}) = p$  e  $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$ .

E, conforme  $n$  aumenta, vimos nos gráficos que:  $\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$ .

# Estimar uma média: Salários

- Quero saber o salário médio das pessoas de uma certa cidade (parâmetro populacional de interesse).
- Posso usar uma amostra e estimar usando a média amostral.
- Quão boa é a estimativa? É precisa?
- Posso pensar no problema de duas formas: Modo 1 e Modo 2.



# Modo 1

- Cidade com  $N$  pessoas.
- Considere

$X_i$  o salário da pessoa  $i$ .

- $\mathbf{X} = (X_1, X_2, \dots, X_N)$ : respostas de toda a população.
- Média populacional:  $\mu = \frac{1}{N} \sum_{i=1}^N X_i$
- Variância populacional:  $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2$



# Modo 1

Temos que,

- $\mu$  = salário médio da população;
- $\sigma^2$  é a variância da população.

Coletamos uma amostra aleatória de tamanho  $n$ .

$\bar{X}$  : média salarial na amostra.

- Quão boa é a estimativa? É precisa?
- Se outra pessoa também coleta uma amostra aleatória de tamanho  $n$  e calcula  $\bar{X}$  teremos o mesmo valor?

# Modo 1 - Exemplo $N = 5$

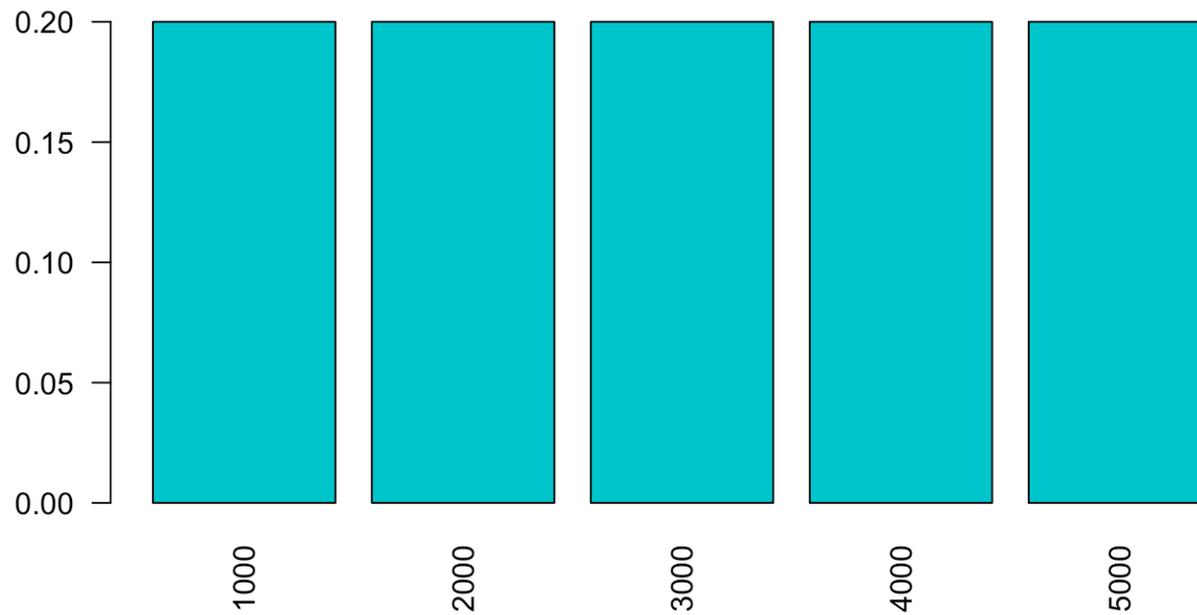
•  $\mathbf{X} = (X_1, X_2, \dots, X_5) = (1000, 2000, 3000, 4000, 5000)$

$$\mu = \frac{\sum_{i=1}^5 X_i}{5} = 3000$$

$$\sigma^2 = \frac{1}{5} \sum_{i=1}^N (X_i - \mu)^2 = 2000000$$

# Modo 1 - Exemplo $N = 5$

Gráfico de barras (proporção) dos dados populacionais:



# Modo 1 - Exemplo $N = 5$ e $n = 2$

- $N^n = 25$  amostras possíveis

$$\mathbf{X} = (X_1, X_2, \dots, X_5) = (1000, 2000, 3000, 4000, 5000)$$

	Pessoa amostrada 1	Pessoa amostrada 2	$\bar{X}$		Pessoa amostrada 1	Pessoa amostrada 2	$\bar{X}$
1	1	1	1000	14	4	3	3500
2	2	1	1500	15	5	3	4000
3	3	1	2000	16	1	4	2500
4	4	1	2500	17	2	4	3000
5	5	1	3000	18	3	4	3500
6	1	2	1500	19	4	4	4000
7	2	2	2000	20	5	4	4500
8	3	2	2500	21	1	5	3000
9	4	2	3000	22	2	5	3500
10	5	2	3500	23	3	5	4000
11	1	3	2000	24	4	5	4500
12	2	3	2500	25	5	5	5000
13	3	3	3000				

# Modo 1 - Exemplo $N = 5$ e $n = 2$

Distribuição amostral de  $\bar{X}$ :

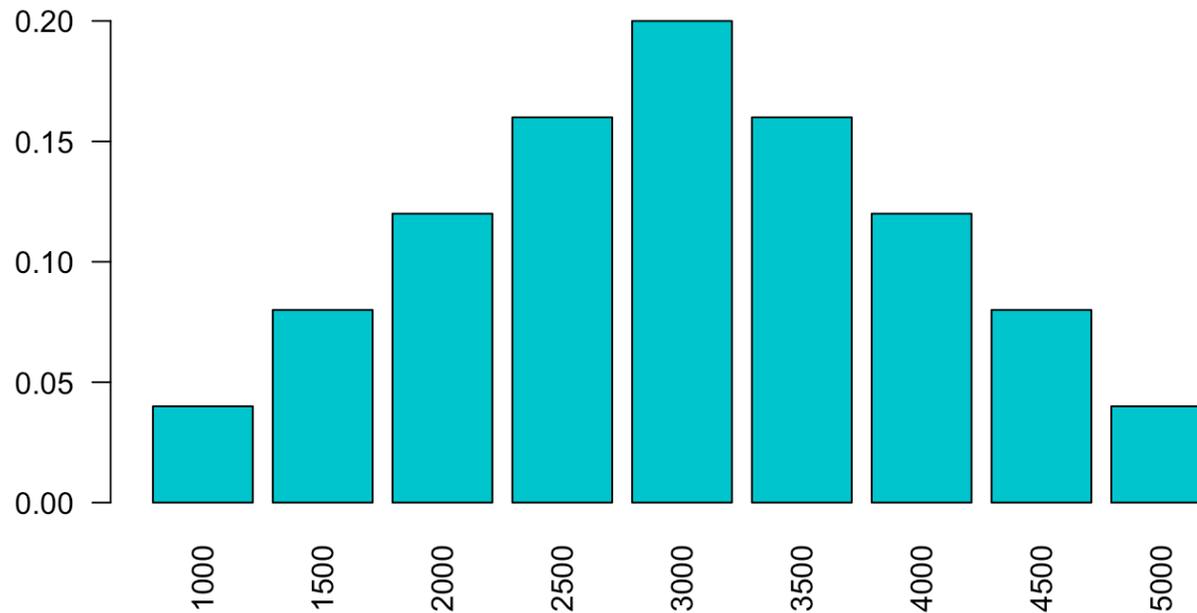
$x$	$P(\bar{X} = x)$
1000	0.04
1500	0.08
2000	0.12
2500	0.16
3000	0.20
3500	0.16
4000	0.12
4500	0.08
5000	0.04

$$E(\bar{X}) = 3000 = \mu$$

$$\begin{aligned} \text{Var}(\bar{X}) &= E[(\bar{X} - \mu)^2] = 10^6 \\ &= \frac{2000000}{2} = \frac{\sigma^2}{n} \end{aligned}$$

# Modo 1 - Exemplo $N = 5$ e $n = 2$

Distribuição amostral de  $\bar{X}$ :



# Modo 1 - Exemplo $N = 5$ e $n = 3$

$N^n = 125$  amostras possíveis.

	Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\bar{X}$		Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\bar{X}$
1	1	1	1	1000.000	14	4	3	1	2666.667
2	2	1	1	1333.333	15	5	3	1	3000.000
3	3	1	1	1666.667	16	1	4	1	2000.000
4	4	1	1	2000.000	17	2	4	1	2333.333
5	5	1	1	2333.333	18	3	4	1	2666.667
6	1	2	1	1333.333	19	4	4	1	3000.000
7	2	2	1	1666.667	20	5	4	1	3333.333
8	3	2	1	2000.000	21	1	5	1	2333.333
9	4	2	1	2333.333	22	2	5	1	2666.667
10	5	2	1	2666.667	23	3	5	1	3000.000
11	1	3	1	1666.667	24	4	5	1	3333.333
12	2	3	1	2000.000	25	5	5	1	3666.667
13	3	3	1	2333.333	26	1	1	2	1333.333

# Modo 1 - Exemplo $N = 5$ e $n = 3$

	Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\bar{X}$		Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\bar{X}$
27	2	1	2	1666.667	42	2	4	2	2666.667
28	3	1	2	2000.000	43	3	4	2	3000.000
29	4	1	2	2333.333	44	4	4	2	3333.333
30	5	1	2	2666.667	45	5	4	2	3666.667
31	1	2	2	1666.667	46	1	5	2	2666.667
32	2	2	2	2000.000	47	2	5	2	3000.000
33	3	2	2	2333.333	48	3	5	2	3333.333
34	4	2	2	2666.667	49	4	5	2	3666.667
35	5	2	2	3000.000	50	5	5	2	4000.000
36	1	3	2	2000.000	51	1	1	3	1666.667
37	2	3	2	2333.333	52	2	1	3	2000.000
38	3	3	2	2666.667	53	3	1	3	2333.333
39	4	3	2	3000.000	54	4	1	3	2666.667
40	5	3	2	3333.333	55	5	1	3	3000.000
41	1	4	2	2333.333	56	1	2	3	2000.000

# Modo 1 - Exemplo $N = 5$ e $n = 3$

	Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\bar{X}$		Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\bar{X}$
57	2	2	3	2333.333	72	2	5	3	3333.333
58	3	2	3	2666.667	73	3	5	3	3666.667
59	4	2	3	3000.000	74	4	5	3	4000.000
60	5	2	3	3333.333	75	5	5	3	4333.333
61	1	3	3	2333.333	76	1	1	4	2000.000
62	2	3	3	2666.667	77	2	1	4	2333.333
63	3	3	3	3000.000	78	3	1	4	2666.667
64	4	3	3	3333.333	79	4	1	4	3000.000
65	5	3	3	3666.667	80	5	1	4	3333.333
66	1	4	3	2666.667	81	1	2	4	2333.333
67	2	4	3	3000.000	82	2	2	4	2666.667
68	3	4	3	3333.333	83	3	2	4	3000.000
69	4	4	3	3666.667	84	4	2	4	3333.333
70	5	4	3	4000.000	85	5	2	4	3666.667
71	1	5	3	3000.000	86	1	3	4	2666.667

# Modo 1 - Exemplo $N = 5$ e $n = 3$

	Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\bar{X}$		Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\bar{X}$
87	2	3	4	3000.000	102	2	1	5	2666.667
88	3	3	4	3333.333	103	3	1	5	3000.000
89	4	3	4	3666.667	104	4	1	5	3333.333
90	5	3	4	4000.000	105	5	1	5	3666.667
91	1	4	4	3000.000	106	1	2	5	2666.667
92	2	4	4	3333.333	107	2	2	5	3000.000
93	3	4	4	3666.667	108	3	2	5	3333.333
94	4	4	4	4000.000	109	4	2	5	3666.667
95	5	4	4	4333.333	110	5	2	5	4000.000
96	1	5	4	3333.333	111	1	3	5	3000.000
97	2	5	4	3666.667	112	2	3	5	3333.333
98	3	5	4	4000.000	113	3	3	5	3666.667
99	4	5	4	4333.333	114	4	3	5	4000.000
100	5	5	4	4666.667	115	5	3	5	4333.333
101	1	1	5	2333.333	116	1	4	5	3333.333

# Modo 1 - Exemplo $N = 5$ e $n = 3$

	Pessoa amostrada 1	Pessoa amostrada 2	Pessoa amostrada 3	$\bar{X}$
117	2	4	5	3666.667
118	3	4	5	4000.000
119	4	4	5	4333.333
120	5	4	5	4666.667
121	1	5	5	3666.667
122	2	5	5	4000.000
123	3	5	5	4333.333
124	4	5	5	4666.667
125	5	5	5	5000.000

# Modo 1 - Exemplo $N = 5$ e $n = 3$

Distribuição amostral de  $\bar{X}$ :

$x$	$P(\bar{X} = x)$	$x$	$P(\bar{X} = x)$
1000	0.008	3333.333	0.144
1333.333	0.024	3666.667	0.12
1666.667	0.048	4000	0.08
2000	0.080	4333.333	0.048
2333.333	0.120	4666.667	0.024
2666.667	0.144	5000	0.008
3000	0.152		

$$E(\bar{X}) = 3000$$

$$= \mu$$

$$Var(\bar{X}) = E[(\bar{X} - \mu)^2]$$

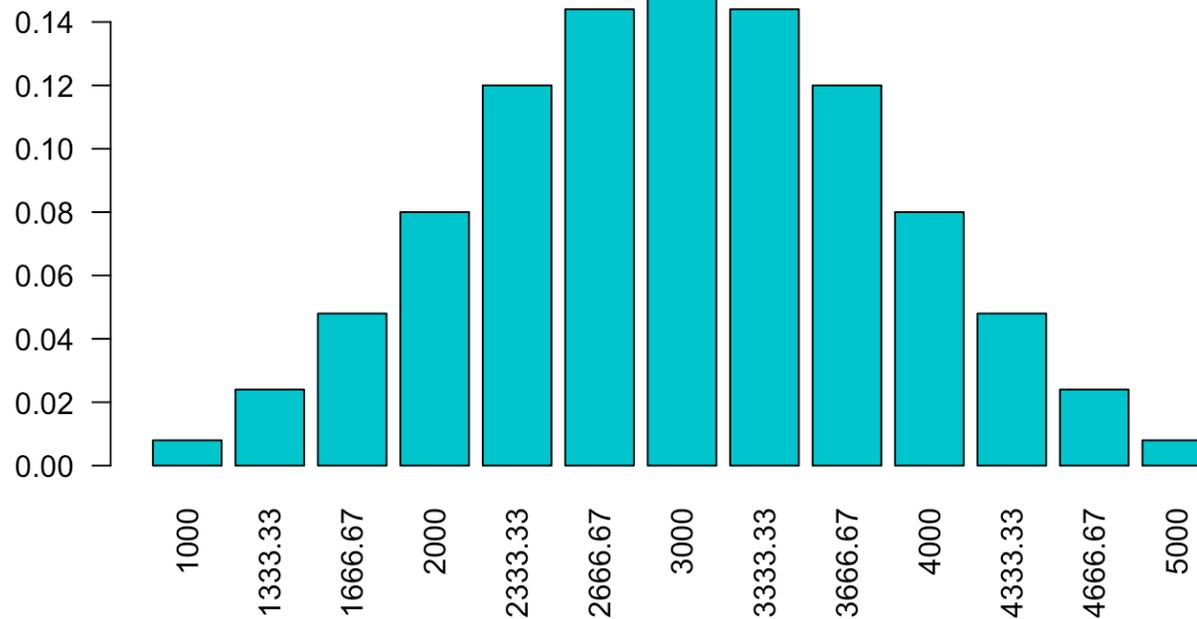
$$= 6.6666668 \times 10^5$$

$$= \frac{2000000}{3}$$

$$= \frac{\sigma^2}{n}$$

# Modo 1 - Exemplo $N = 5$ e $n = 3$

Distribuição amostral de  $\bar{X}$ :



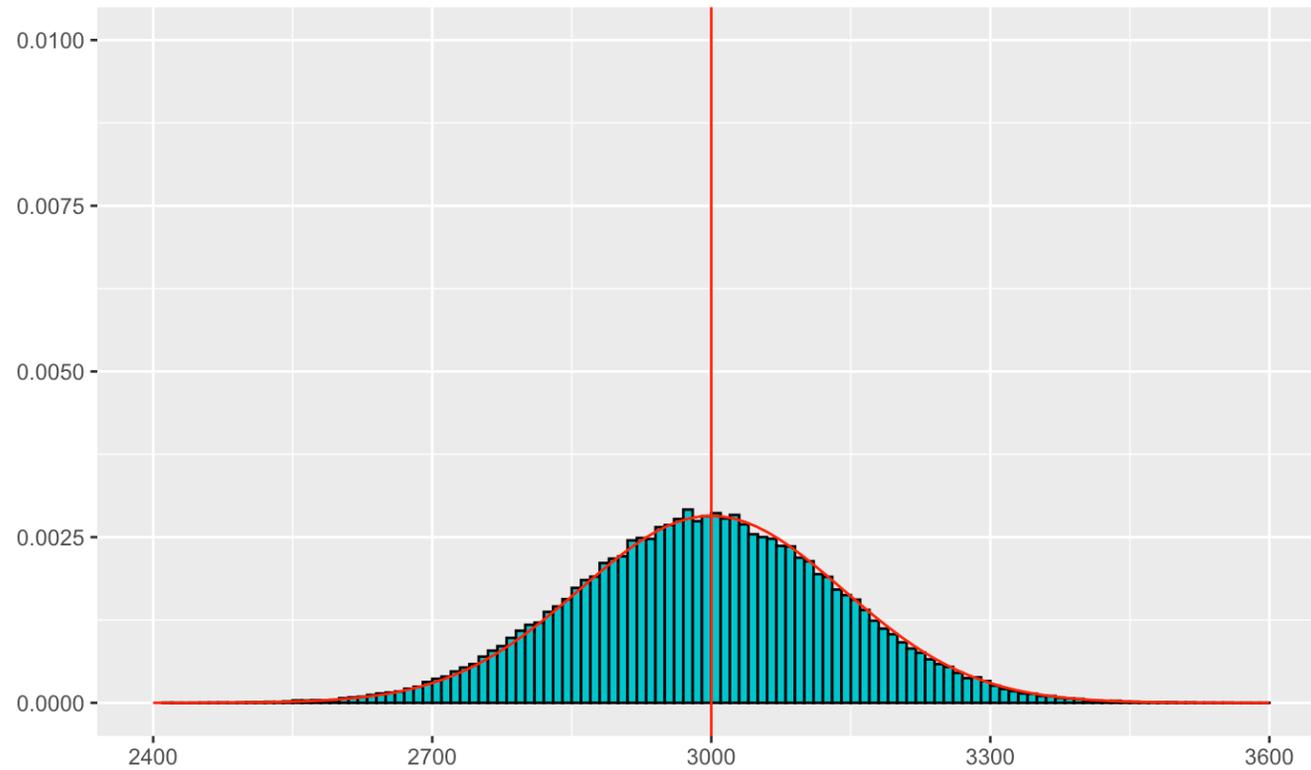
# Modo 1

- $\mathbf{X} = (X_1, \dots, X_N)$  é fixo
- Amostra aleatória de tamanho  $n$
- $\bar{X}$  é v.a.
- $E(\bar{X}) = \mu$
- $Var(\bar{X}) = \frac{\sigma^2}{n}$



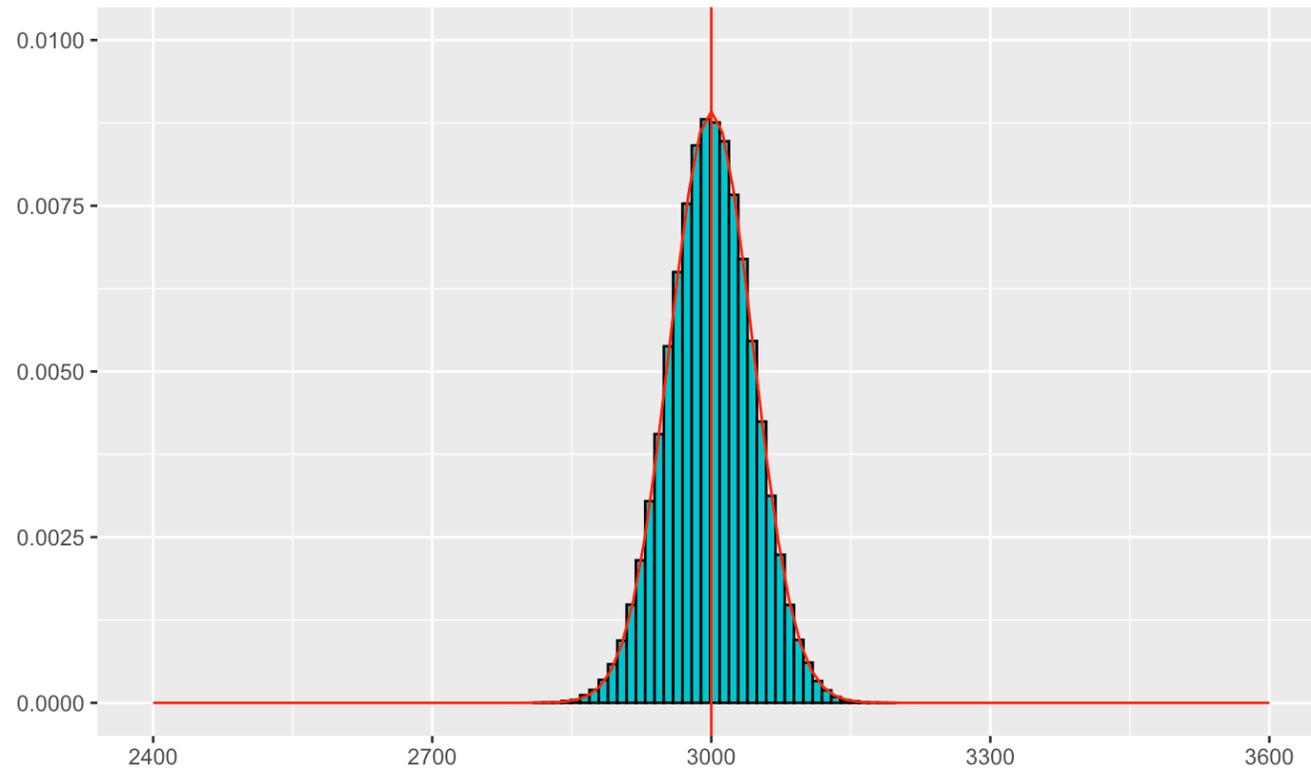
# Modo 1 - Exemplo $N = 1000000$ e $n = 100$

- $\mu = 3000$ . Distribuição amostral de  $\bar{X}$ :



# Modo 1 - Exemplo $N = 1000000$ e $n = 1000$

- $\mu = 3000$ . Distribuição amostral de  $\bar{X}$ :



## Modo 2

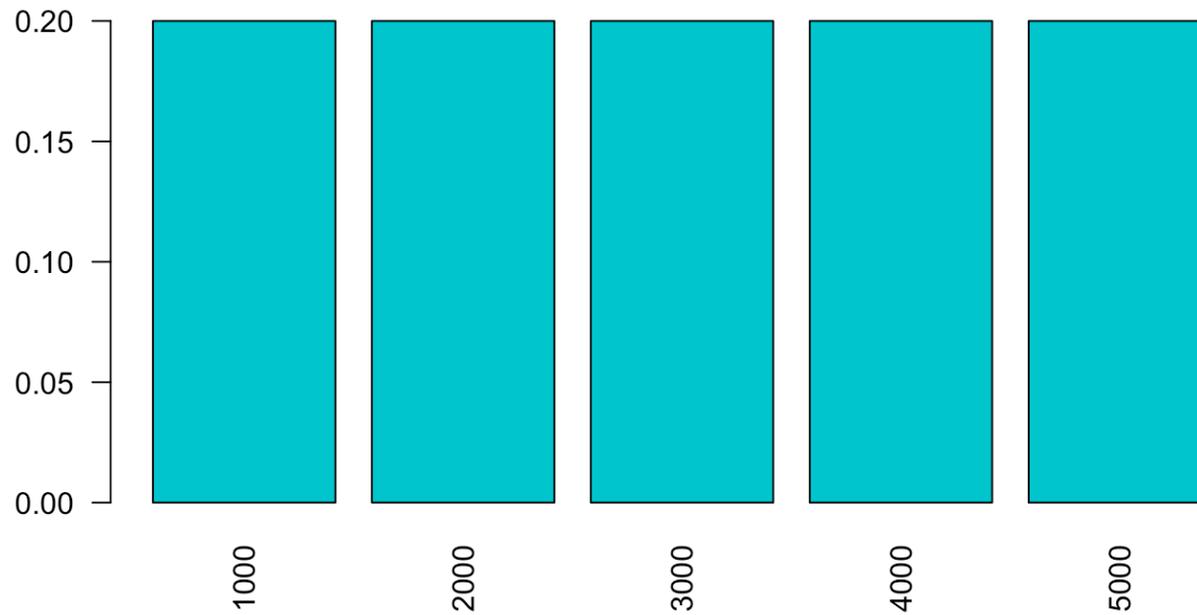
Suponha que o salário de uma pessoa possa ser representado por uma **variável aleatória** uniforme discreta assumindo os valores 1000, 2000, 3000, 4000 ou 5000.

$$\mu = \mathbb{E}(X) = \frac{1000 + 2000 + 3000 + 4000 + 5000}{5} = 3000$$

$$\begin{aligned}\sigma^2 = \text{Var}(X) &= \frac{1}{5}[(1000 - 3000)^2 + (2000 - 3000)^2 + (3000 - 3000)^2 \\ &+ (4000 - 3000)^2 + (5000 - 3000)^2] \\ &= 2000000\end{aligned}$$

# Modo 2

Distribuição da variável  $X$  (do salário de cada indivíduo da população):



# Modo 2 - Exemplo $n = 2$

Temos que

$$\mathbb{E}(\bar{X}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) = E(X) = \mu = 3000$$

$$\text{Var}(\bar{X}) = \frac{1}{n} \sum_{i=1}^n \text{Var}(X_i) = \frac{\text{Var}(X)}{n} = \frac{\sigma^2}{n} = 1000000$$

(propriedades de linearidade da esperança e variância (a.a.))

# Resultado

Seja  $X$  uma v.a. com média  $\mu$  e variância  $\sigma^2$  e  $X_1, \dots, X_n$  uma amostra aleatória de  $X$ .

A média amostral

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

tem as seguintes propriedades:

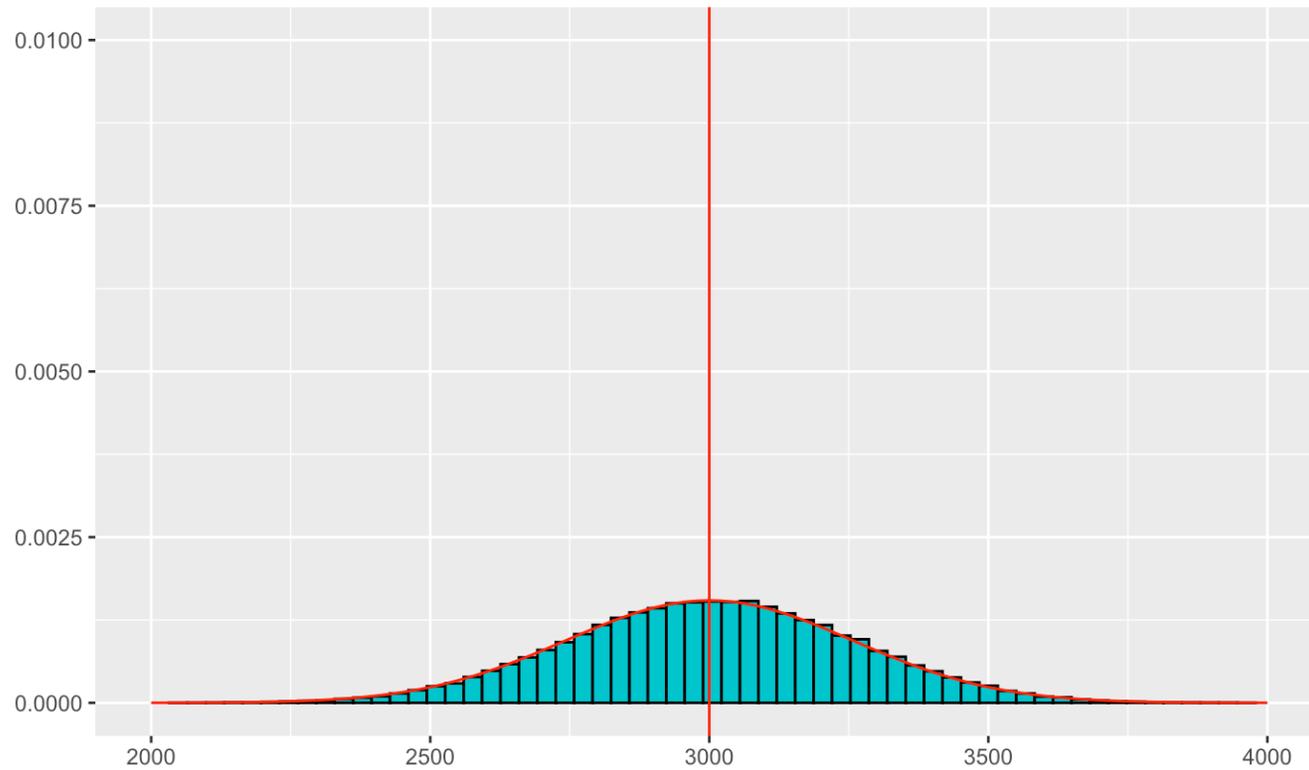
$$\mathbb{E}(\bar{X}_n) = \mu \quad \text{e} \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

(propriedade de linearidade da esperança e da variância, esta última em caso de independência)

Ou seja, embora  $\mu$  seja desconhecido, sabemos que o valor esperado da média amostral é  $\mu$ . Além disso, conforme o tamanho amostral aumenta, a imprecisão da média amostral para estimar  $\mu$  fica cada vez menor, pois  $\text{Var}(\bar{X}) = \sigma^2/n$  é inversamente proporcional ao tamanho amostral  $n$ .

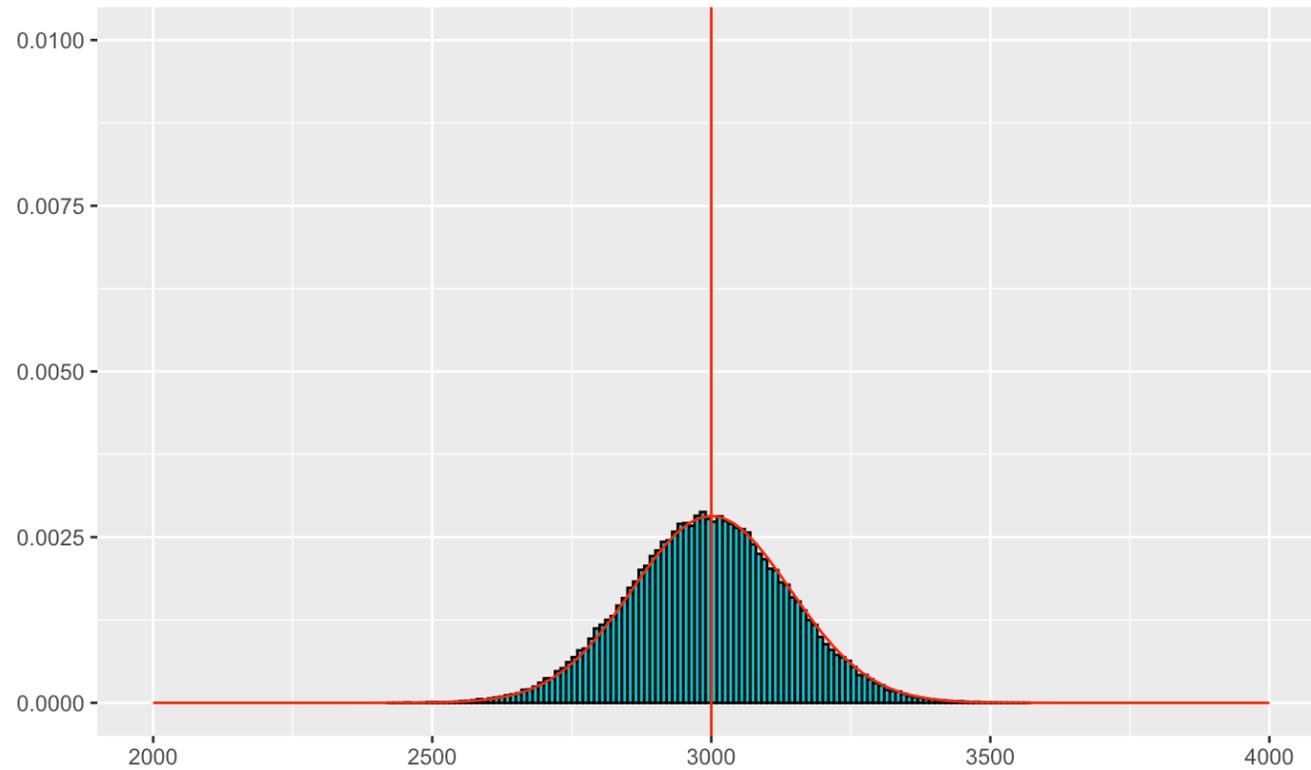
# Modo 2 - Exemplo $n = 30$

•  $\mu = 3000$



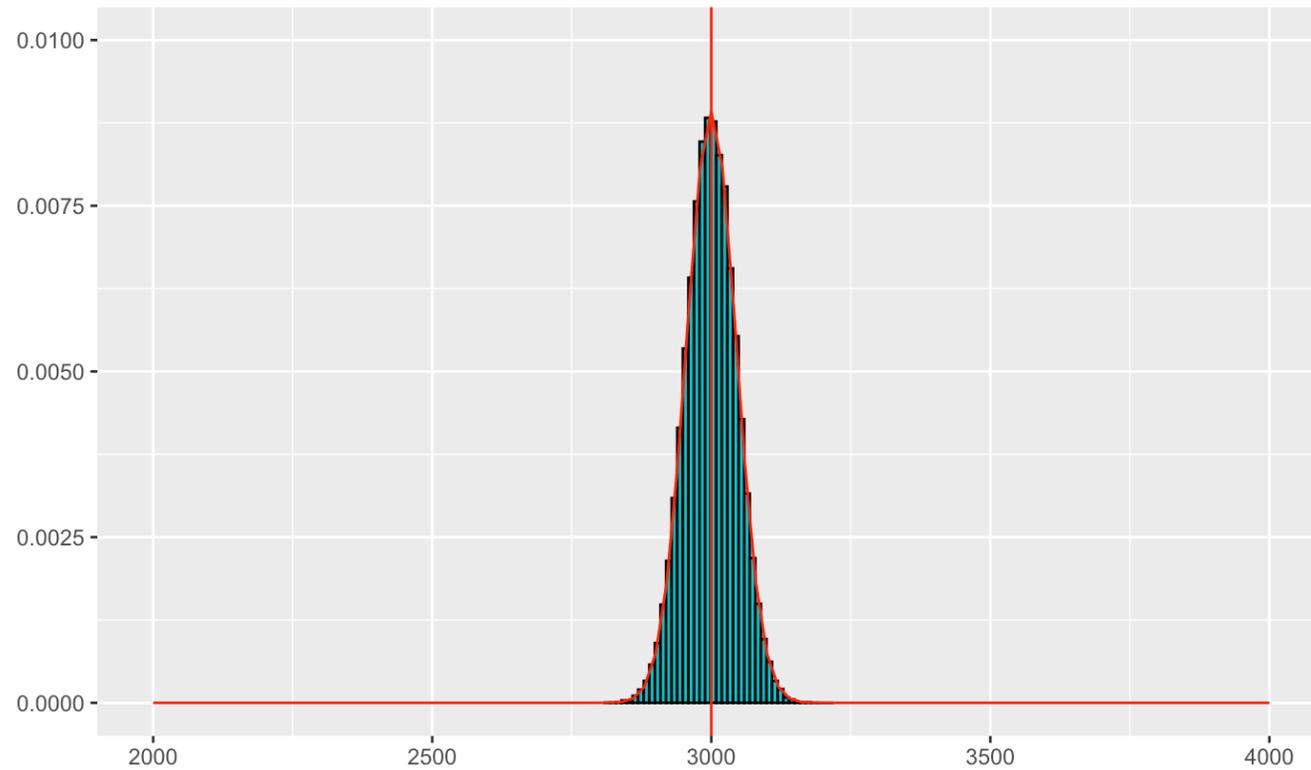
# Modo 2 - Exemplo $n = 100$

•  $\mu = 3000$



# Modo 2 - Exemplo $n = 1000$

•  $\mu = 3000$



# Resultados

Temos uma população com média (proporção)  $\mu$  ( $p$ ) e variância  $\sigma^2$  desconhecida.

Retira-se uma amostra aleatória de tamanho  $n$  e calcula-se a média (ou proporção) amostral  $\bar{X}$  (ou  $\hat{p}$ ) para estimar o parâmetro populacional desconhecido  $\mu$  (ou  $p$ ).

Temos as propriedades:

$$E(\bar{X}) = \mu \quad \text{e} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n};$$

$$E(\hat{p}) = p \quad \text{e} \quad \text{Var}(\hat{p}) = \frac{p(1-p)}{n}.$$

E, conforme  $n$  aumenta, pelos gráficos, *parece* que a distribuição amostral de  $\bar{X}$  e  $\hat{p}$  se aproxima da normal:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{e} \quad \hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

Para uma amostra aleatória de tamanho  $n$  a partir de uma população:

- com média  $\mu$  e variância  $\sigma^2$

$$\bar{X}: \mathbb{E}(\bar{X}) = \mu \text{ e } \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

$$\text{Erro padrão: } EP(\bar{X}) = \sqrt{\text{Var}(\bar{X})} = \sigma/\sqrt{n}.$$

- com proporção populacional  $p$

$$\hat{p}: \mathbb{E}(\hat{p}) = p \text{ e } \text{Var}(\hat{p}) = \frac{p(1-p)}{n}.$$

$$\text{Erro padrão: } EP(\hat{p}) = \sqrt{\text{Var}(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}.$$

No exemplos, vimos também a distribuição amostral de  $\bar{X}$  ou  $\hat{p}$ , mas isso só foi possível porque tínhamos informação de todos os valores possíveis na população.

Os exemplos anteriores foram casos hipotéticos apenas para ilustrar como  $\bar{X}$  e  $\hat{p}$  se comportam quando realizamos a amostragem.

Na prática, não teremos informações suficientes para de fato descrevermos a distribuição amostral exata de  $\bar{X}$  e  $\hat{p}$  (se tivermos, nem é preciso fazer amostragem!)

# Teorema do Limite Central

## Resultado

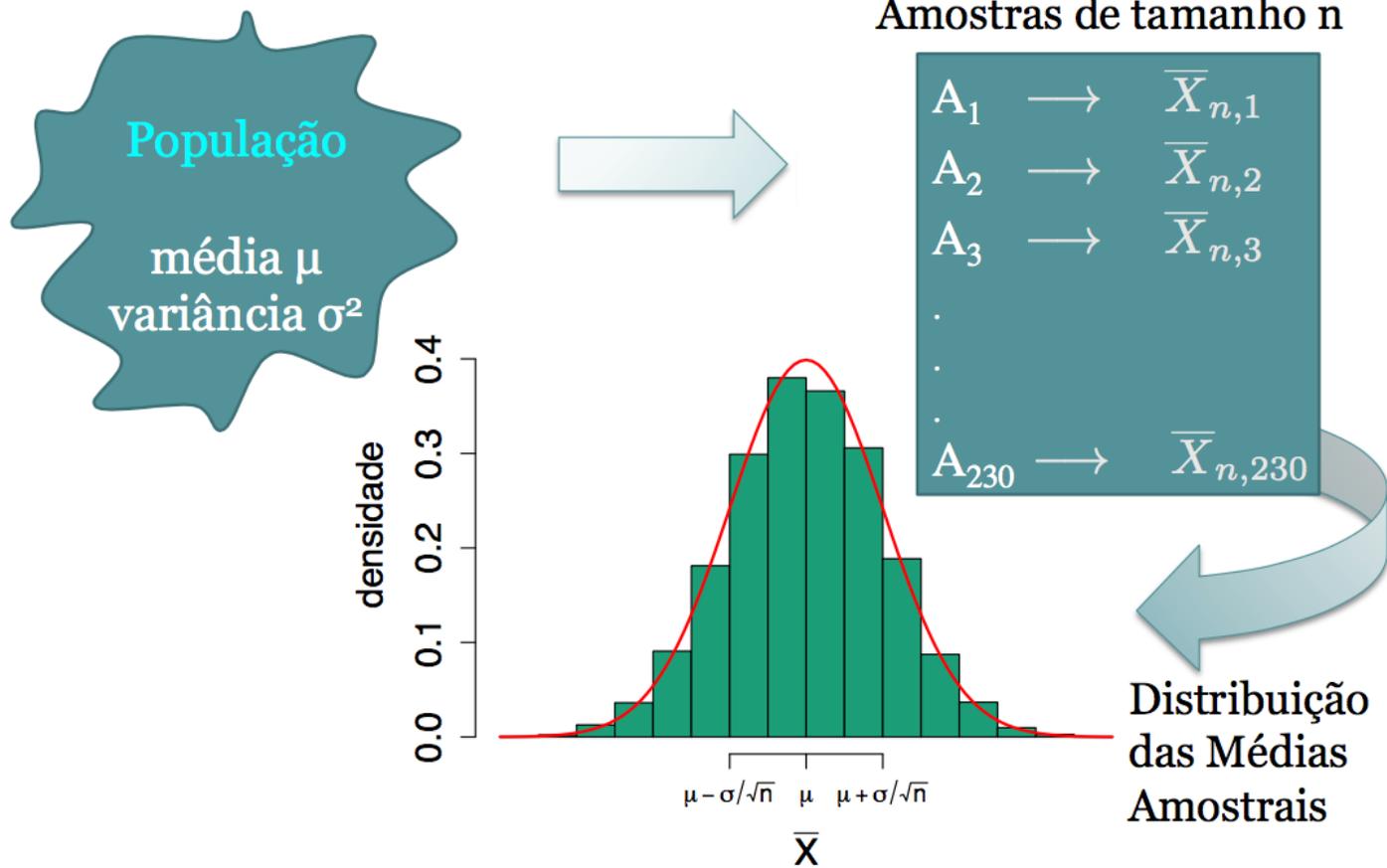
- Para uma amostra aleatória  $X_1, \dots, X_n$  coletada de uma população com média  $\mu$  e variância  $\sigma^2$ , a distribuição amostral de  $\bar{X}$  aproxima-se de uma **distribuição Normal** de média  $\mu$  e variância  $\frac{\sigma^2}{n}$ , quando  $n$  for suficientemente grande:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

*Obs:* O resultado vale para  $\hat{p}$ , com  $\mu = p$  e  $\sigma^2 = p(1 - p)$ .

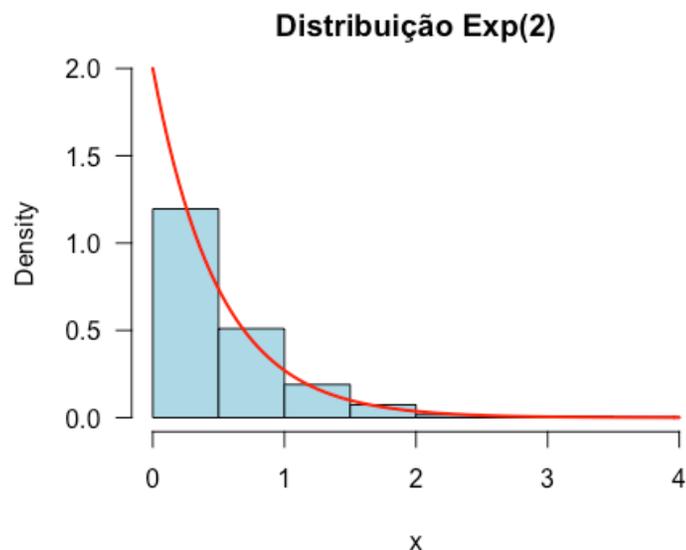
# Teorema do Limite Central



# Exemplo - Transistor

Suponha que  $X$  denota o tempo de vida de um transistor (em horas) e seu comportamento pode ser representado por uma distribuição Exponencial, tal que  $X \sim \text{Exp}(2)$ , ou seja:

$$f_X(x) = 2e^{-2x}, \quad \text{para } x \geq 0.$$



Sabemos que:

$$\mathbb{E}(X) = \frac{1}{2}$$

$$\text{Var}(X) = \frac{1}{4}$$

# Exemplo - Transistor

Os tempos de vida de 100 transistores escolhidos ao acaso são coletados e a média dos tempos é calculada, denotada por  $\bar{X}_{100}$ .

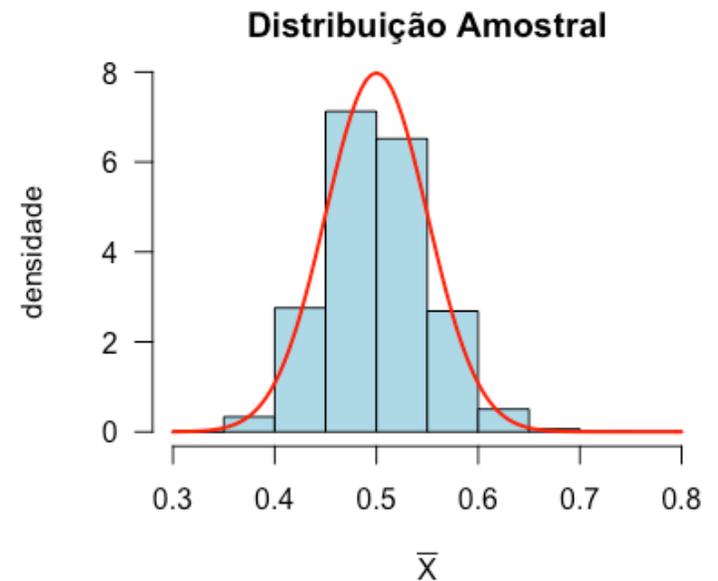
Desejamos estudar a variável aleatória  $\bar{X}_{100}$ .

Sabemos que:

$$\mathbb{E}(\bar{X}_{100}) = \frac{1}{2} \text{ e } \text{Var}(\bar{X}_{100}) = \frac{1/4}{100} = \frac{1}{400}$$

Então, pelo TLC:

$$\bar{X}_{100} \sim N\left(\frac{1}{2}, \frac{1}{400}\right).$$



# Exemplo - Lançamento de dados

Considere  $X$  = resultado obtido no lançamento de um dado honesto.

$x$	1	2	3	4	5	6
$p(x) = P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

---

$$\mathbb{E}(X) = \frac{1}{6} \times (1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3.5$$

$$\text{Var}(X) = \frac{1}{6} [(1 - 3.5)^2 + (2 - 3.5)^2 + \dots + (6 - 3.5)^2] = \frac{17.5}{6} = 2.92$$

Além disso, considere

- $X_i$ : resultado do  $i$ -ésimo lançamento de um dado honesto.
- $X_i$  tem distribuição uniforme discreta, com  
$$\mu = \mathbb{E}(X_i) = 3.5 \quad \text{e} \quad \sigma^2 = \text{Var}(X_i) = 2.92.$$

# Exemplo - Lançamento de dados

Se temos uma amostra aleatória de tamanho  $n$ :  $X_1, X_2, \dots, X_n$ , pelo TLC sabemos que a distribuição amostral de  $\bar{X}$  é Normal  $\left(3.5, \frac{2.92}{n}\right)$  para  $n$  grande.

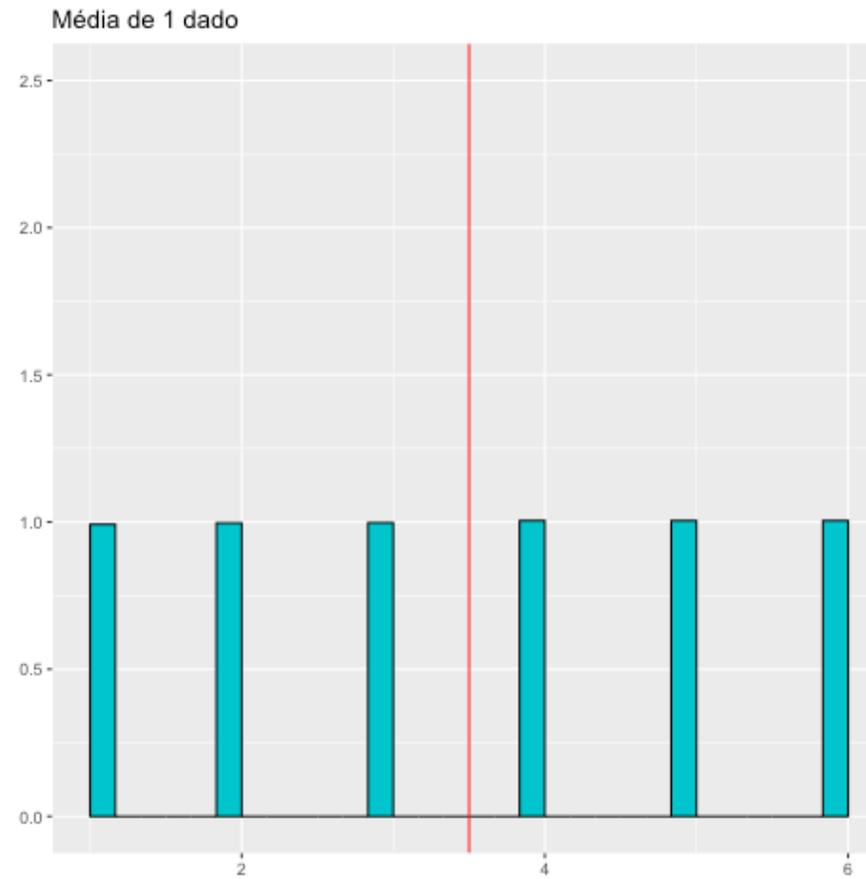
O primeiro histograma a seguir mostra o resultado de 100000 repetições do seguinte experimento: observar o resultado do lançamento de 1 dado. Repare que é muito próximo de uma distribuição uniforme discreta (chance 1/6 para cada resultado), que é a distribuição de  $X$ .

O segundo histograma mostra o resultado de 100000 repetições do seguinte experimento: observar a média do lançamento de 2 dados.

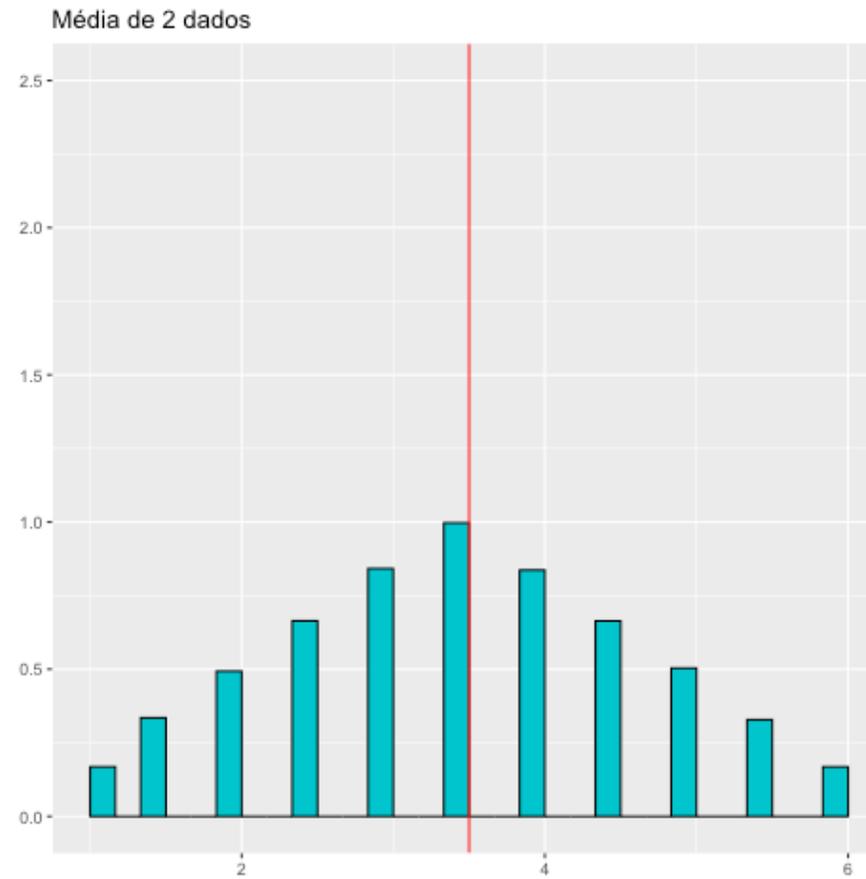
O último histograma mostra o resultado de 100000 repetições do seguinte experimento: observar a média do lançamento de 100 dados.

Repare que conforme o número de dados lançados (tamanho amostral) aumenta, a distribuição da média amostral se aproxima da distribuição normal com média 3.5 e variância cada vez menor ( $2.92/n$ ).

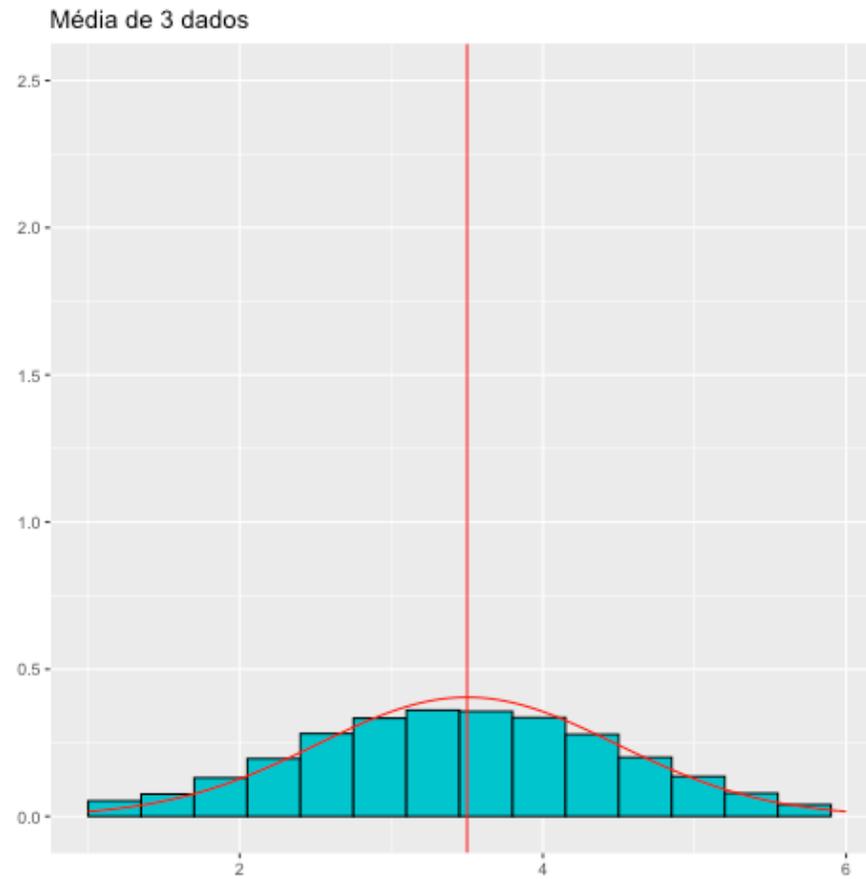
# Exemplo - Lançamento de dados



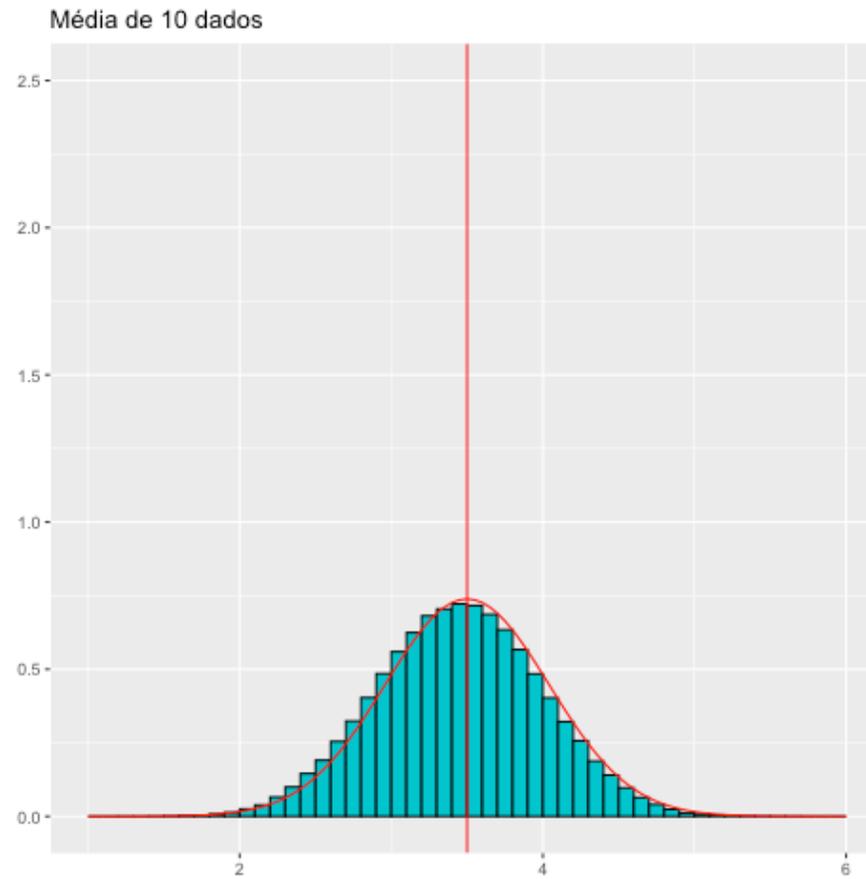
# Exemplo - Lançamento de dados



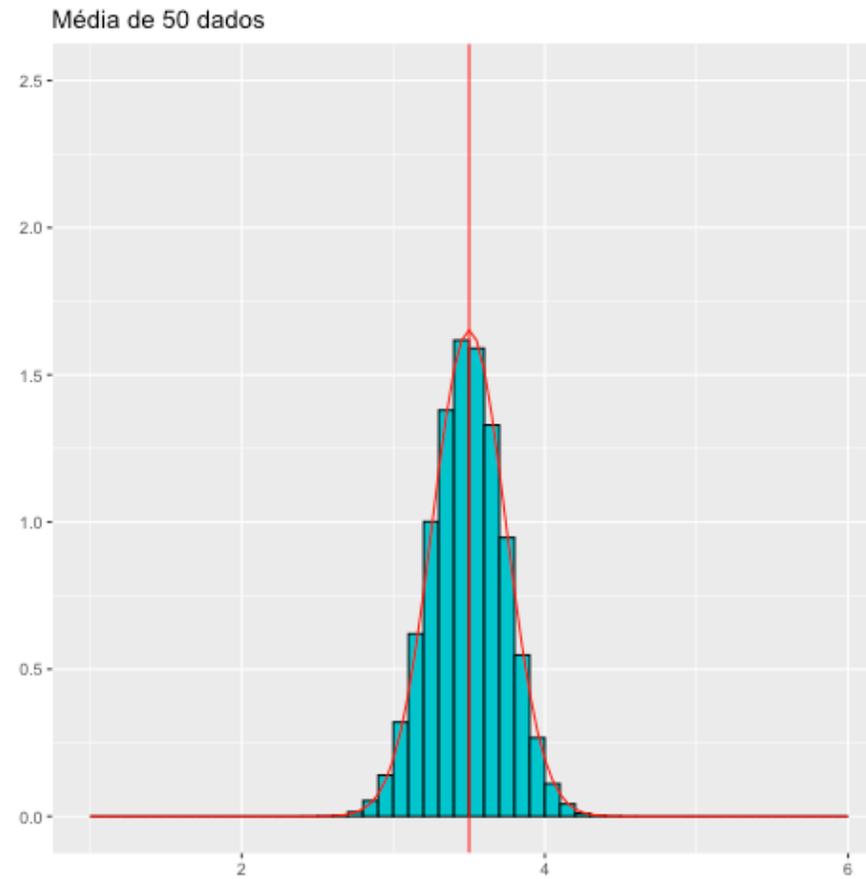
# Exemplo - Lançamento de dados



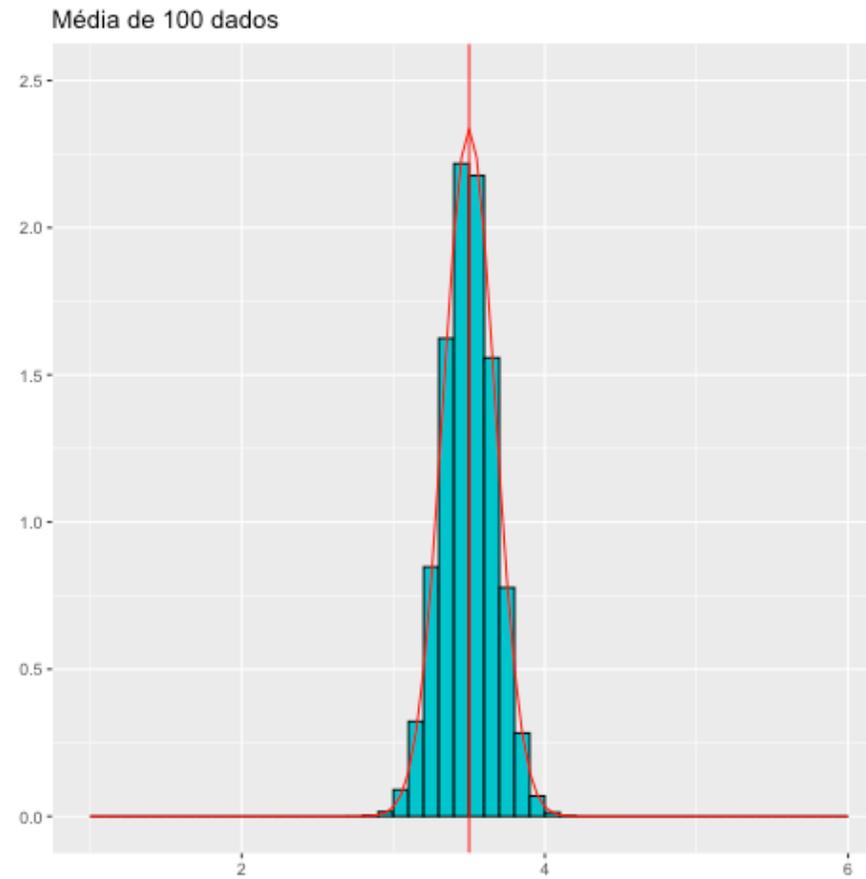
# Exemplo - Lançamento de dados



# Exemplo - Lançamento de dados



# Exemplo - Lançamento de dados



# Teorema do Limite Central (TLC)

Você pode verificar o comportamento de  $\bar{X}$  para várias distribuições de  $X$ :

[TLC para proporções](#)

[TLC para médias](#)

# Aproximação da Binomial pela Normal

$$\text{Se } \hat{p} = \frac{\sum_{i=1}^n X_i}{n} = \frac{S_n}{n} \implies S_n = n\hat{p}.$$

Quando  $n$  é grande o suficiente:  $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$ .

Nesse caso, qual a distribuição de  $S_n$ ?

- Vimos que  $S_n = X_1 + \dots + X_n \sim \text{Bin}(n, p)$ .
- Pelas propriedades da distribuição Normal:

$$S_n = n\hat{p} \sim N(np, np(1-p)).$$

- Portanto, quando  $n$  é grande,  $\text{Bin}(n, p) \approx N(np, np(1-p))$ .

# Fundamentos de Inferência

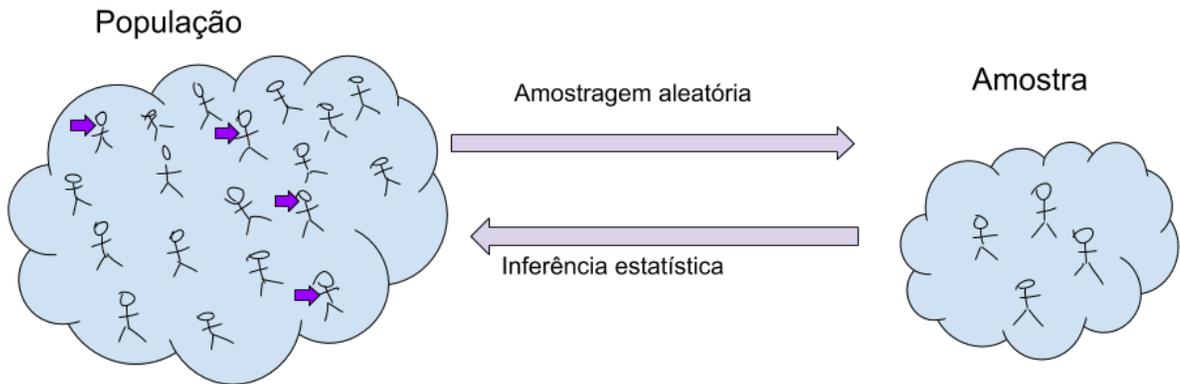
# Introdução

Um dos principais objetivos da Estatística é tirar conclusões a partir dos dados.

Dados em geral consistem de uma amostra de elementos de uma população de interesse.

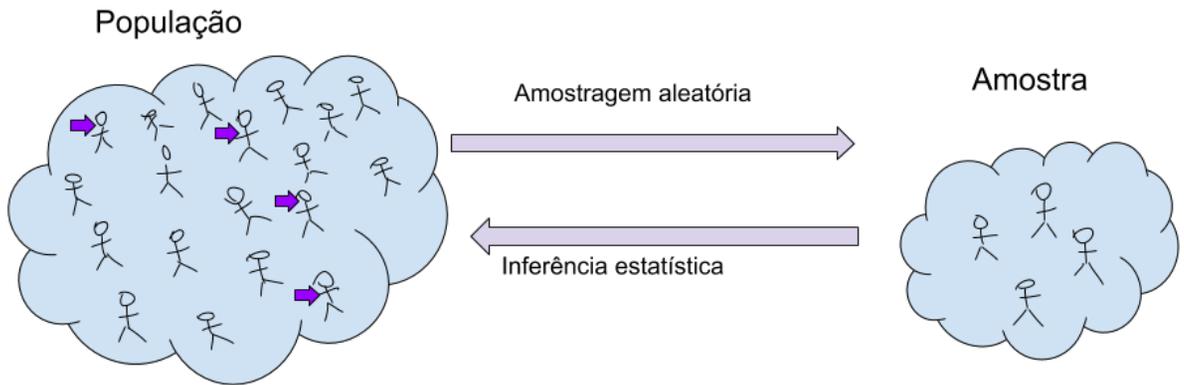
Usar a amostra para tirar conclusões sobre a população.

Quão confiável será utilizar a informação obtida apenas de uma amostra para concluir algo sobre a população?



# Introdução

**População:** todos os elementos ou resultados de um problema que está sendo estudado.

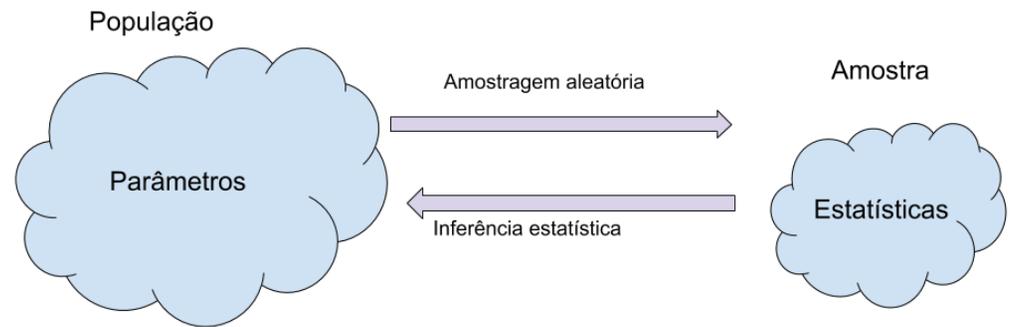


**Amostra:** subconjunto da população de interesse.

# Inferência Estatística

**Variável:** Característica numérica do resultado de um experimento.

**Parâmetros:** Característica numérica (desconhecida) da distribuição dos elementos da população.

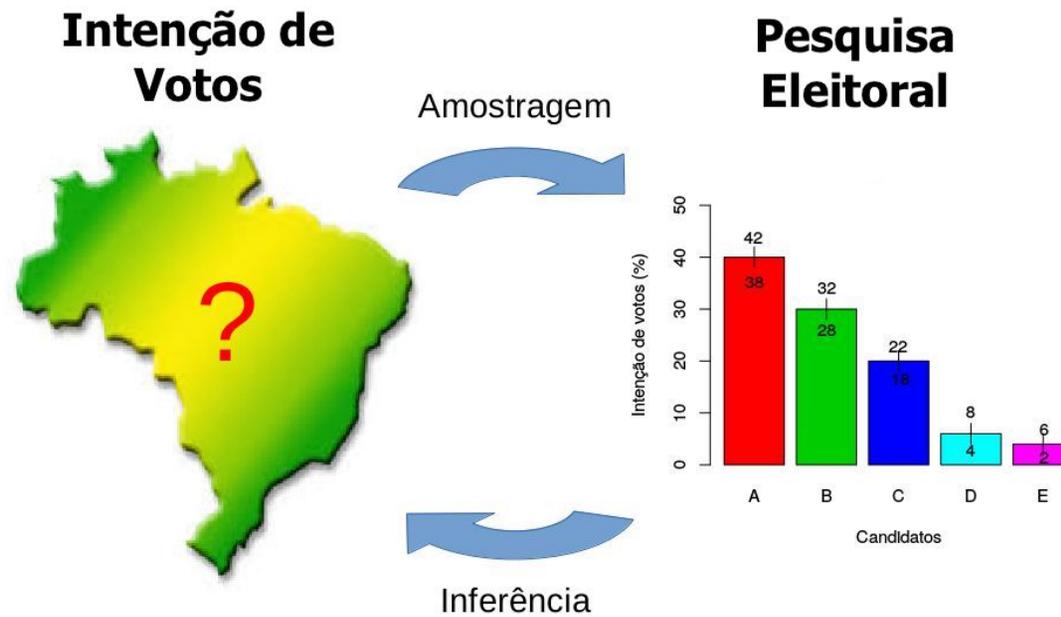


**Estimador/Estatística:** Função da amostra, construída com a finalidade de representar, ou estimar um parâmetro de interesse na população.

**Estimativa:** Valor numérico que um estimador assume para uma dada amostra.

**Erro amostral:** é a diferença entre um estimador e o parâmetro que se quer estimar.

# Inferência Estatística



# Estatística

Seja  $X_1, \dots, X_n$  uma amostra e

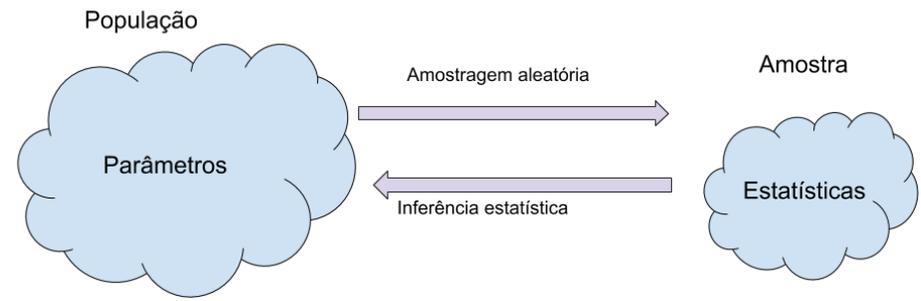
$$T = f(X_1, \dots, X_n)$$

é uma estatística.

Exemplos:

- $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n}(X_1 + \dots + X_n)$
- $X_{(1)} = \min\{X_1, \dots, X_n\}$  ou  $X_{(n)} = \max\{X_1, \dots, X_n\}$
- $X_{(i)}$  é o  $i$ -ésimo valor da amostra ordenada

Note que uma estatística é uma função que em uma determinada amostra assume um valor específico (estimativa).



# Estatística

Para que serve uma estatística?

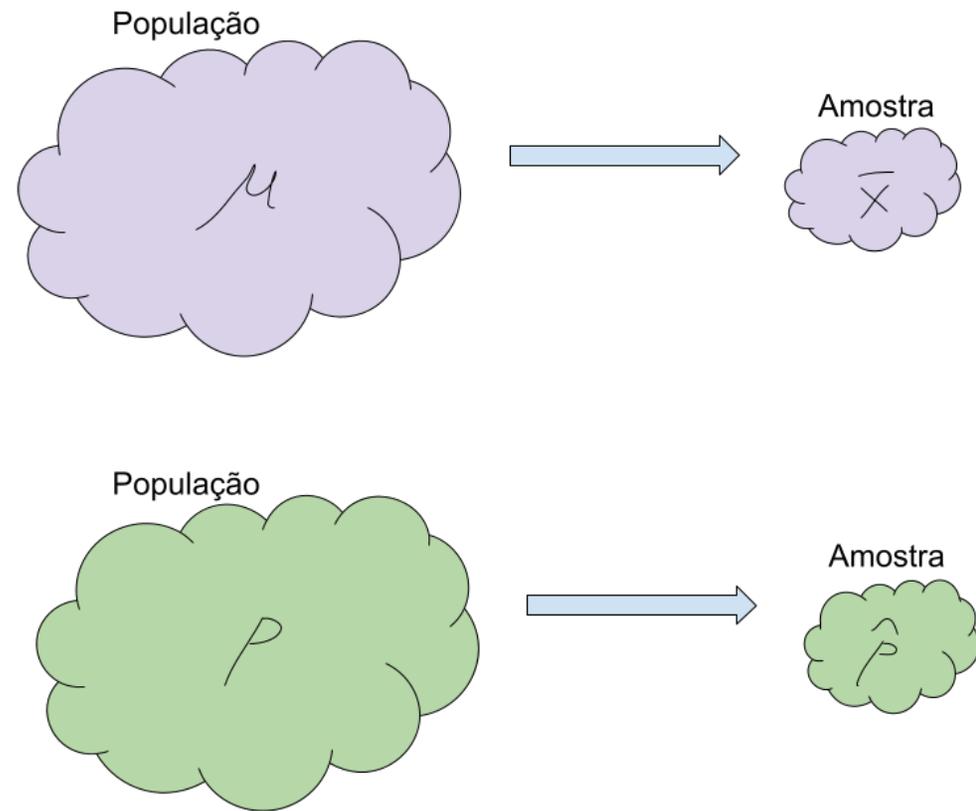
Para “estimar” características de uma população.

**População:**

- Média  $\mu$
- Proporção  $p$

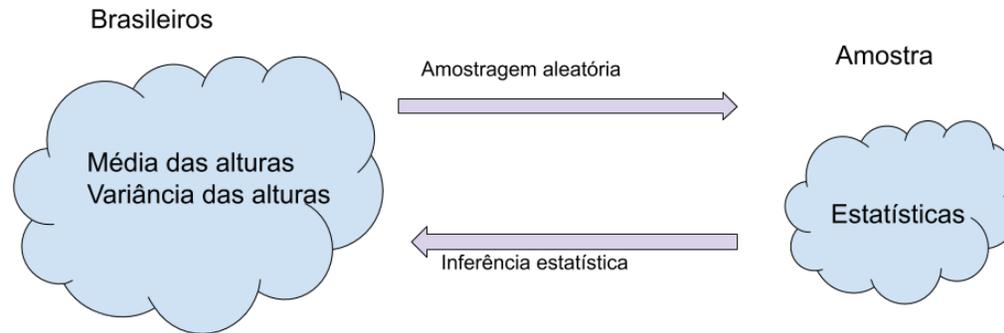
**Amostra:**

- Média Amostral  
$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$
- Proporção Amostral  
$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$$



# Exemplo

Temos interesse em saber a média e a variância da altura dos brasileiros:  $\mu$  e  $\sigma^2$ .



**Solução 1:** Medir a altura de todos os brasileiros.

**Solução 2:** Selecionar de forma aleatória alguns brasileiros (amostra), analisá-la e inferir propriedades para toda a população.

# Parâmetro

- Cada quantidade de interesse (como  $\mu$  e  $\sigma^2$  no exemplo anterior) é chamada de parâmetro da população.
- Para apresentar uma estimativa de um parâmetro ( $\hat{\mu}$  e  $\hat{\sigma}^2$ ), devemos escolher uma estatística ( $T$ ).
- Note que da maneira que o plano amostral foi executado (amostra aleatória), a estatística  $T$  é uma variável aleatória, visto que cada vez que executarmos o plano amostral poderemos obter resultados diversos.
- Portanto, a estatística  $T$  possui uma distribuição de probabilidade, chamada de **distribuição amostral de T**.

# Leituras

- [Ross](#): capítulo 7.
- [OpenIntro](#): seção 4.1.
- Magalhães: capítulo 7.



Slides produzidos pelos professores:

- Samara Kiihl
- Tatiana Benaglia
- Larissa Matos
- Benilton Carvalho

[Fonte da imagem](#)