# Bioestatística

## INFERÊNCIA ESTATÍSTICA

Silvia Shimakura

# AMOSTRAS E POPULAÇÕES

- Inferências sobre populações são geralmente feitas a partir de informações obtidas de amostras.
- Válido se a amostra for representativa da população.
- Para assegurar que não há viés sistemático selecionamos aleatoriamente membros da população.
- Amostra aleatória independente:
  - Todos os elementos da população têm iguais chances de serem selecionados.
  - Todas as combinações possíveis de um dado número de elementos têm a mesma chance de serem selecionados.

# Estimação

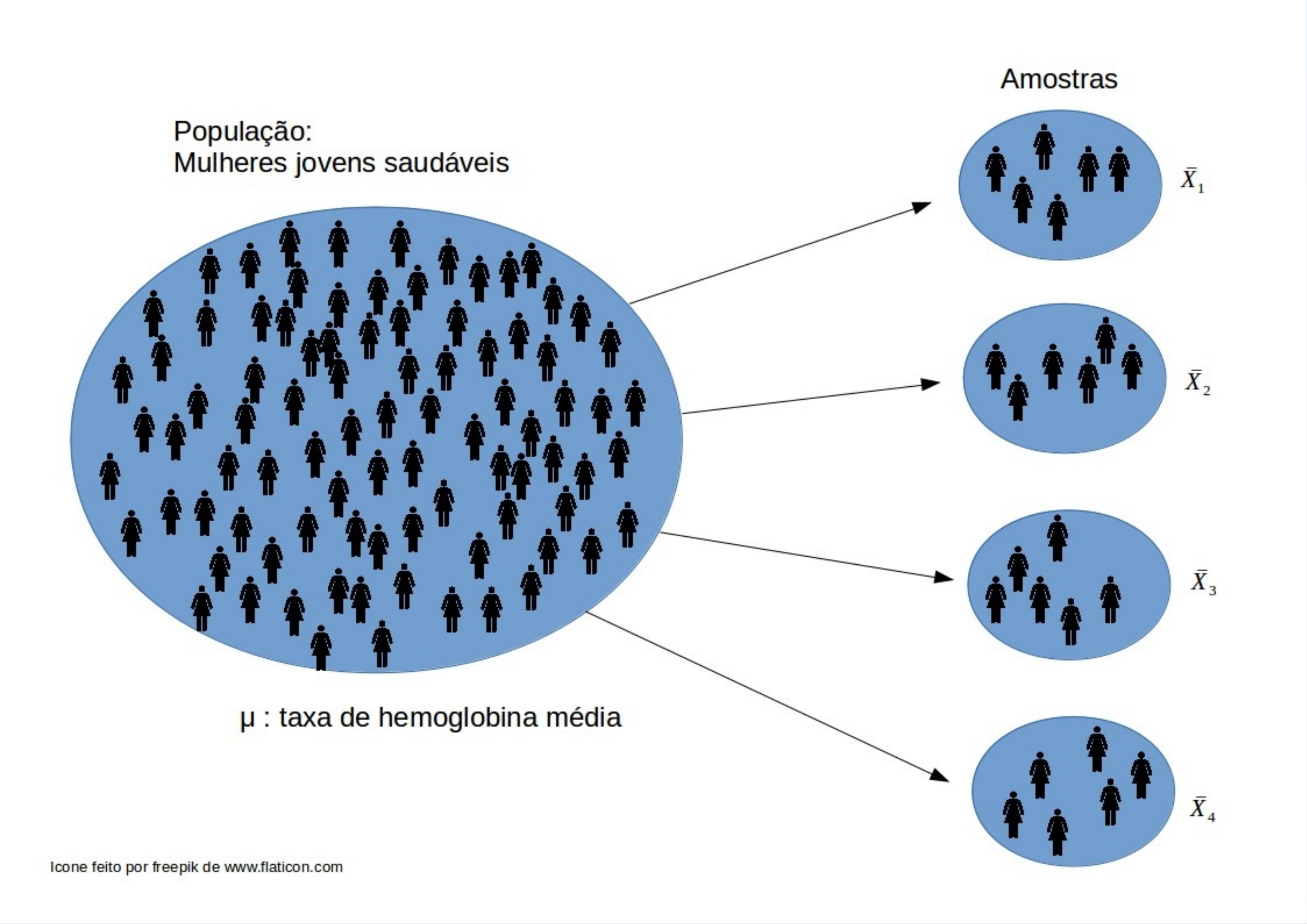
- Amostras são usadas para estimar quantidades desconhecidas da população.
- **EXEMPLO:** prevalência de uma doença, efeito de uma intervenção, diferença entre grupos
- É importante saber qual é a variação destas estimativas de amostra para amostra.

# Estimação

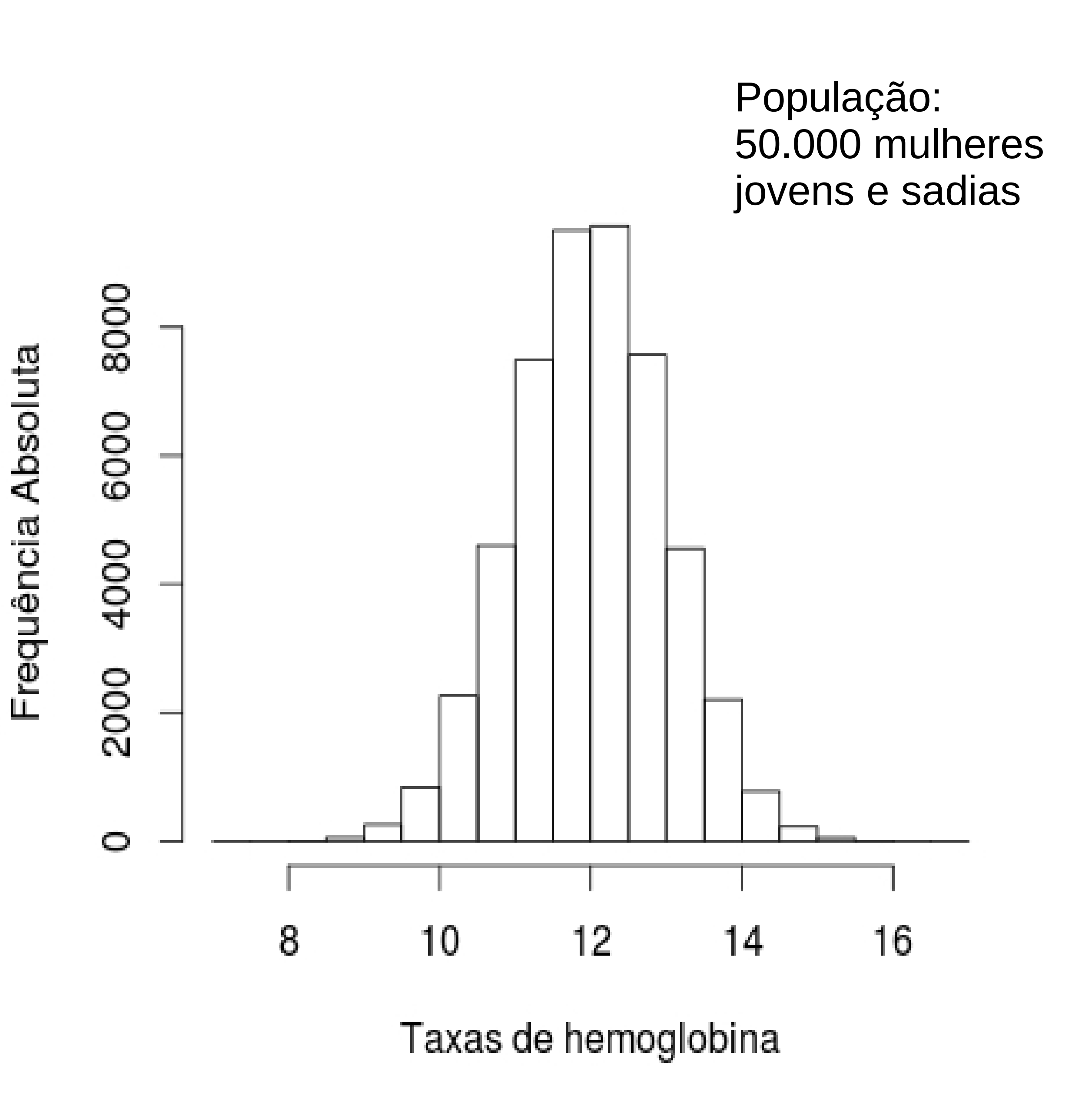
- Teoria de probabilidades permite usar amostras para estimar quantidades de populações, e determinar a precisão destas estimativas.

# Estimação de uma média

- O que acontece quando retiramos diversas amostras de uma população e estimamos a média da população usando as médias amostrais?



# Distribuição das taxas de hemoglobina




- Média=12
- Desvio-padrão=1
- Na prática a média e o desvio-padrão são desconhecidos!!!
- Censo inviável ou impossível.
- Conclusões são baseadas numa amostra.

# Amostragem 1

- Uma amostra de tamanho n=6 é selecionada da população de taxas de hemoglobina.

| | | | | | | |
|---|---|---|---|---|---|---|
| Amostra 1 | 11,75 | 11,26 | 11,80 | 12,95 | 11,62 | 10,86 |
| Média 1 | 11,71 | | | | | |

# Amostragem 2

- Selecionando-se outras 6 mulheres...temos um resultado diferente...

| | | | | | | |
|---|---|---|---|---|---|---|
| Amostra 1 | 11,75 | 11,26 | 11,80 | 12,95 | 11,62 | 10,86 |
| Média 1 | 11,71 | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Amostra 2 | 11,43 | 12,60 | 10,86 | 10,93 | 12,24 | 13,76 |
| Média 2 | 11,97 | | | | | |

- A média amostral varia de uma amostra para outra!

# PERGUNTAS

- É possível estimar a média populacional e determinar a precisão da estimativa?

- Existe um comportamento sistemático das médias amostrais?

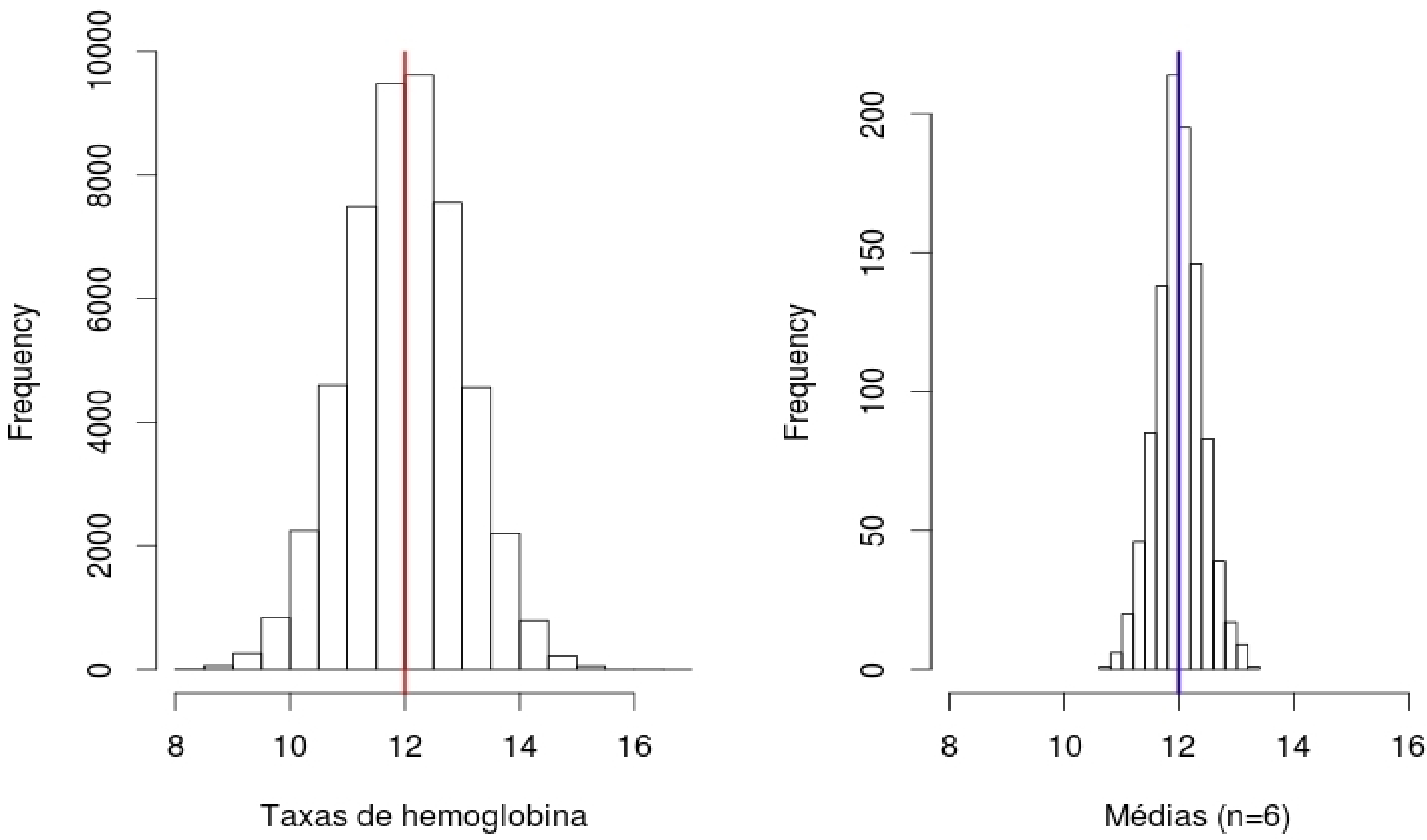
# RESPOSTA

- Vamos tentar responder as perguntas com um exercício de simulação.
- Selecionamos 1000 amostras de 6 mulheres e calculamos as médias amostrais.

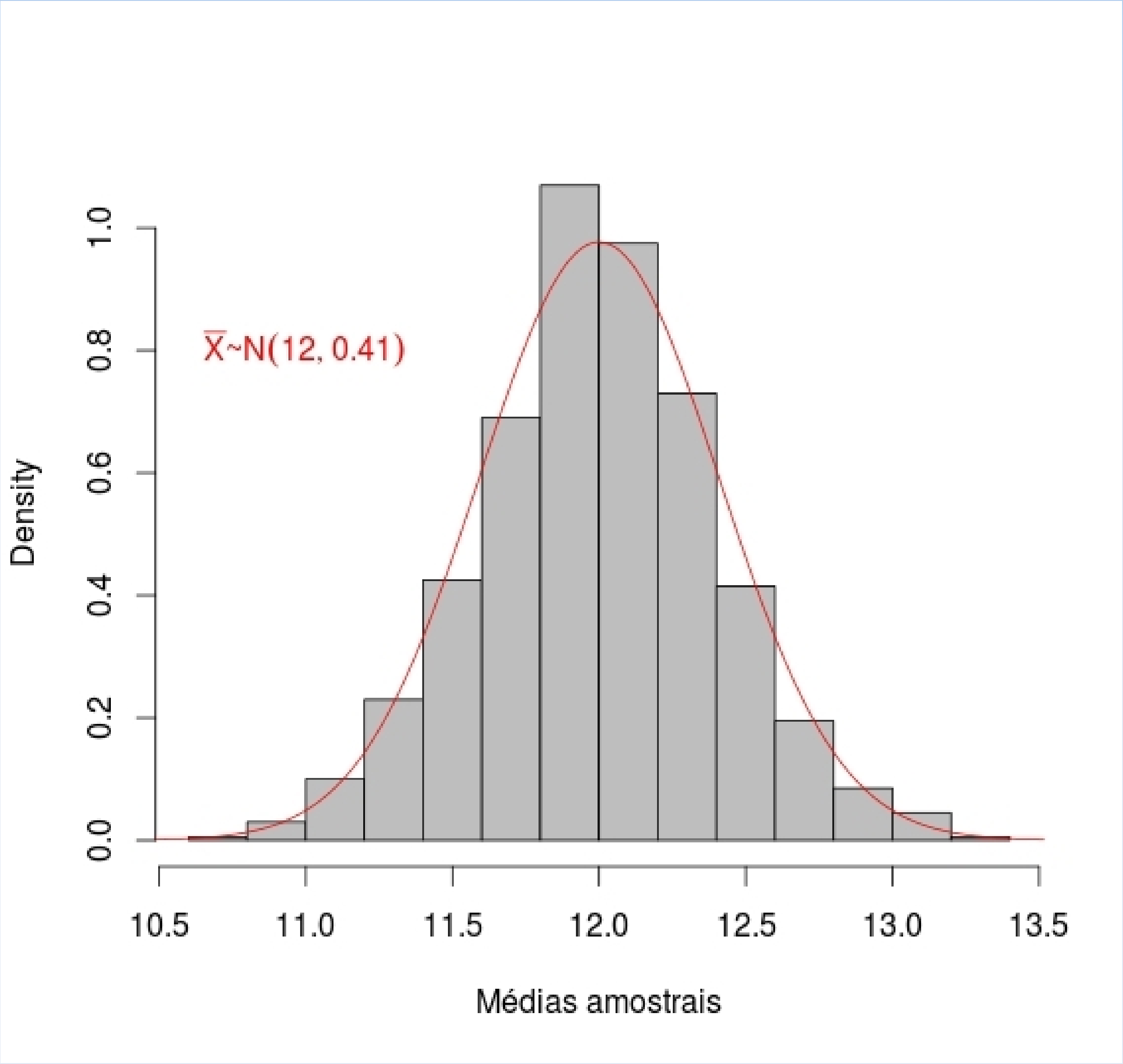
| Amostra | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 11,78 | 11,48 | 10,91 | 11,35 | 11,95 | 10,95 | 12,32 | 12,18 | 12,41 | 10,58 |
| | 11,46 | 10,71 | 11,11 | 10,42 | 10,14 | 11,35 | 12,25 | 12,20 | 14,35 | 12,74 |
| | 13,41 | 13,06 | 11,31 | 13,57 | 12,01 | 11,83 | 11,33 | 11,50 | 12,29 | 10,42 |
| | 12,33 | 11,11 | 12,66 | 11,47 | 13,05 | 9,81 | 11,50 | 11,21 | 12,31 | 12,59 |
| | 11,02 | 12,69 | 11,33 | 11,75 | 12,07 | 12,72 | 12,29 | 10,05 | 13,49 | 12,21 |
| | 12,19 | 11,62 | 11,42 | 12,93 | 13,12 | 12,84 | 10,42 | 13,61 | 11,12 | 11,47 |
| Média | 12,03 | 11,78 | 11,46 | 11,92 | 12,06 | 11,58 | 11,69 | 11,79 | 12,66 | 11,67 |

- As médias amostrais $\left(\bar{X}\right)$ variam de acordo com alguma distribuição de probabilidade conhecida?

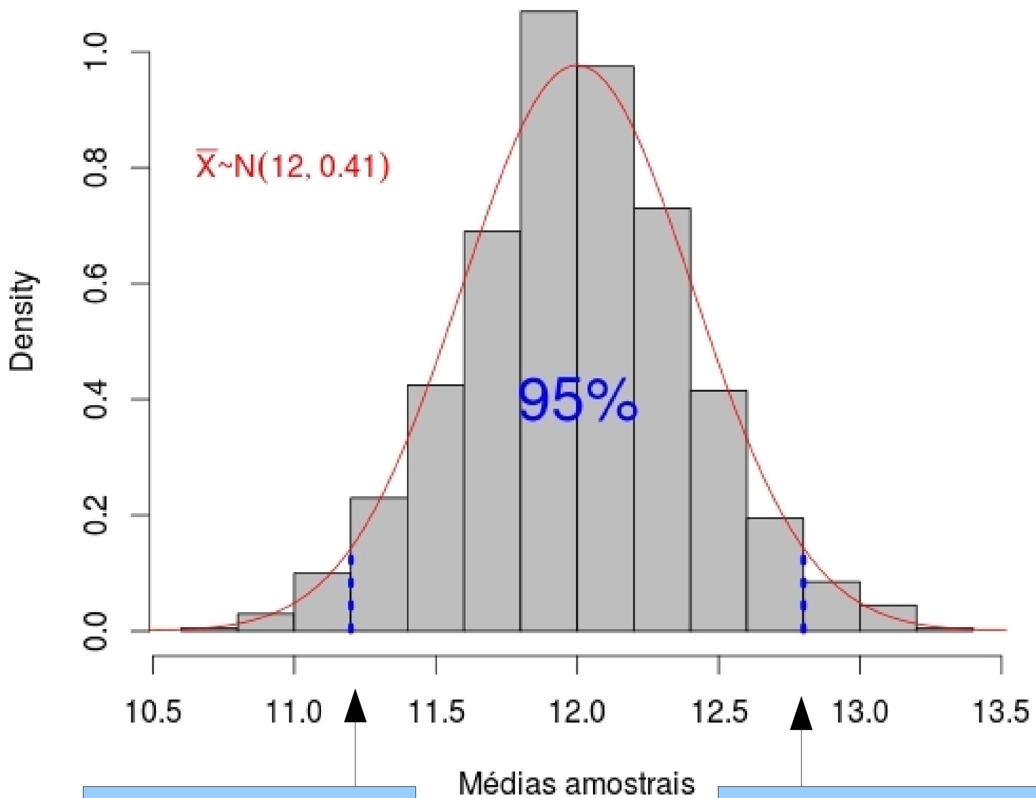
# Distribuição população x média



# Teorema Central do Limite

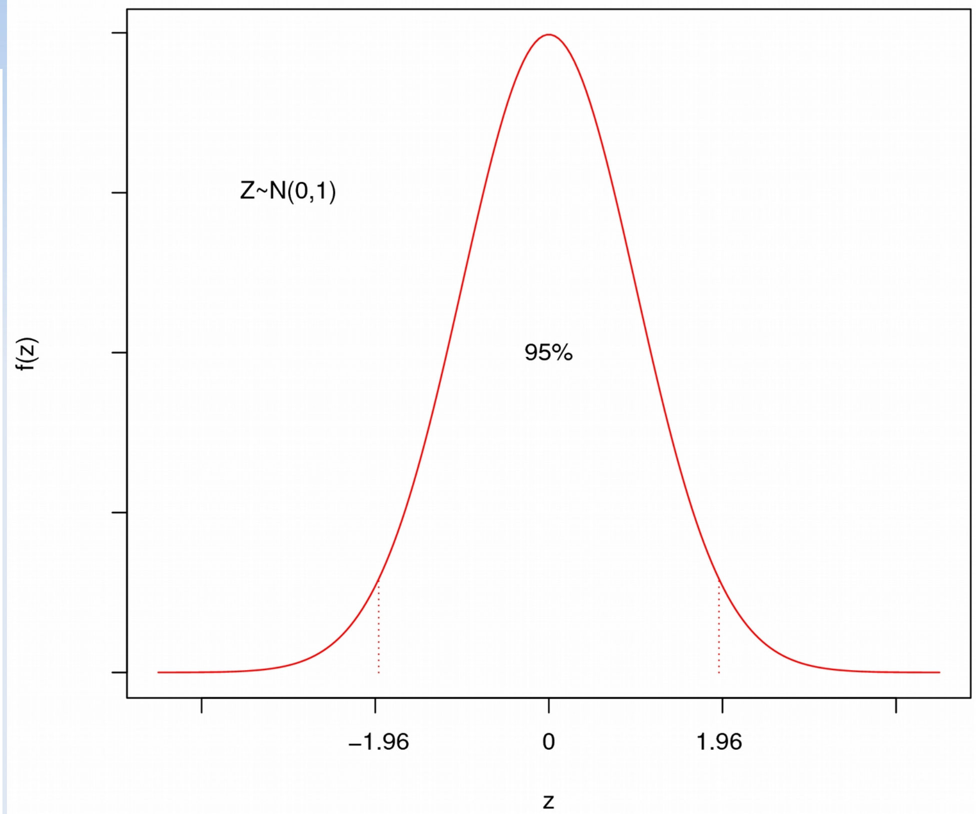


# Teorema Central do Limite



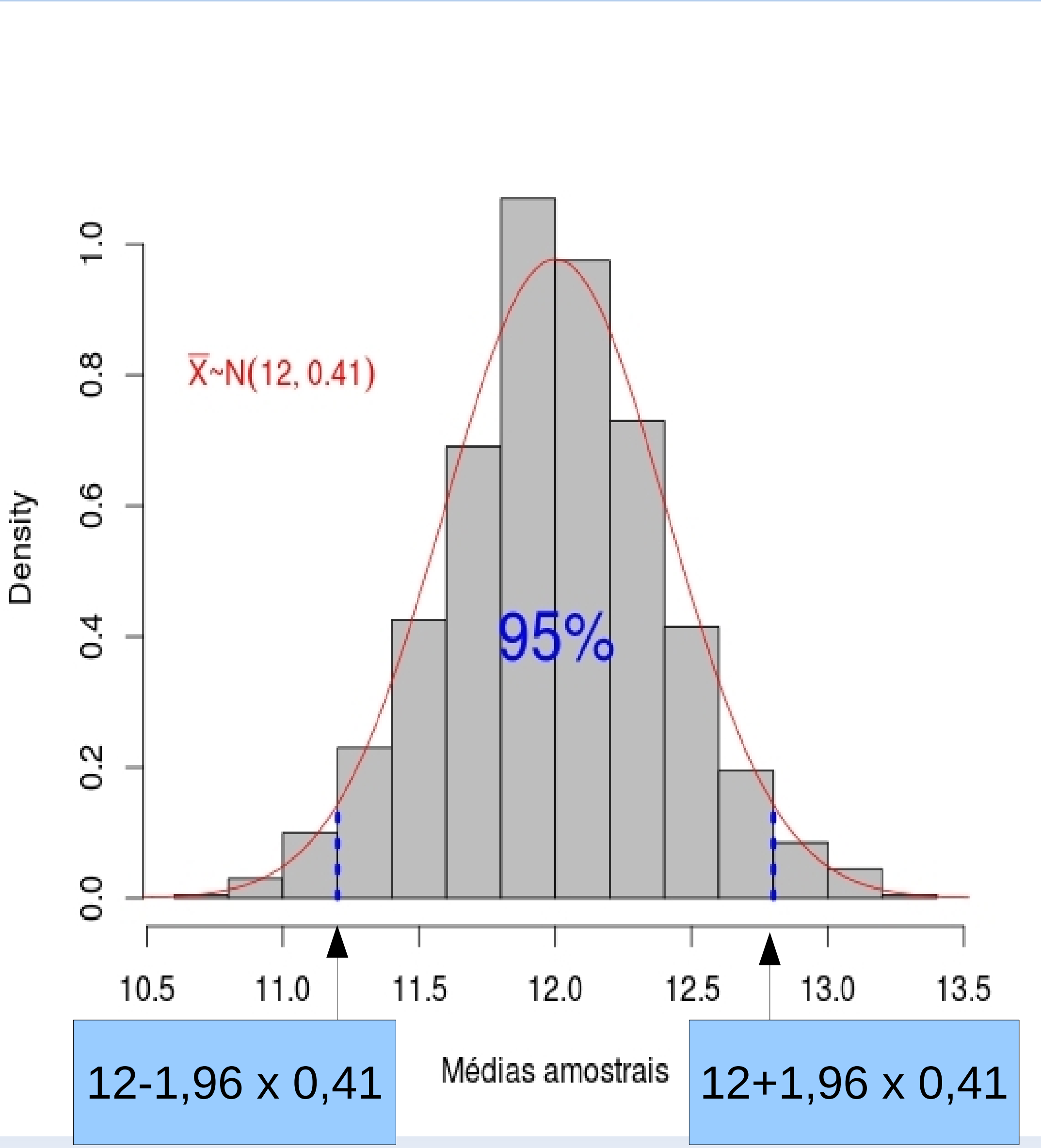

12-1,96 x 0,41

12+1,96 x 0,41




$$\frac{\bar{X}-12}{0{,}41} = Z \sim N(0{,}1)$$

# Consequência do TCL




- 95% das médias amostrais estão entre (12±1,96x0,41)

$$P(12-1,96\,x\,0,41 \leqslant \bar{X} \leqslant 12+1,96\,x\,0,41)=0,95$$

$$\Downarrow$$

$$P(\bar{X}-1,96\times 0,41 \leqslant 12 \leqslant \bar{X}+1,96\times 0,41)=0,95$$

- 95% dos intervalos ( $\bar{X}$ ±1,96x0,41) cobrem 12

- 95% das médias amostrais estão entre (12±1,96x0,41)

$$P\left(12-1{,}96\times0{,}41\leqslant\bar{X}\leqslant12+1{,}96\times0{,}41\right)=0{,}95$$

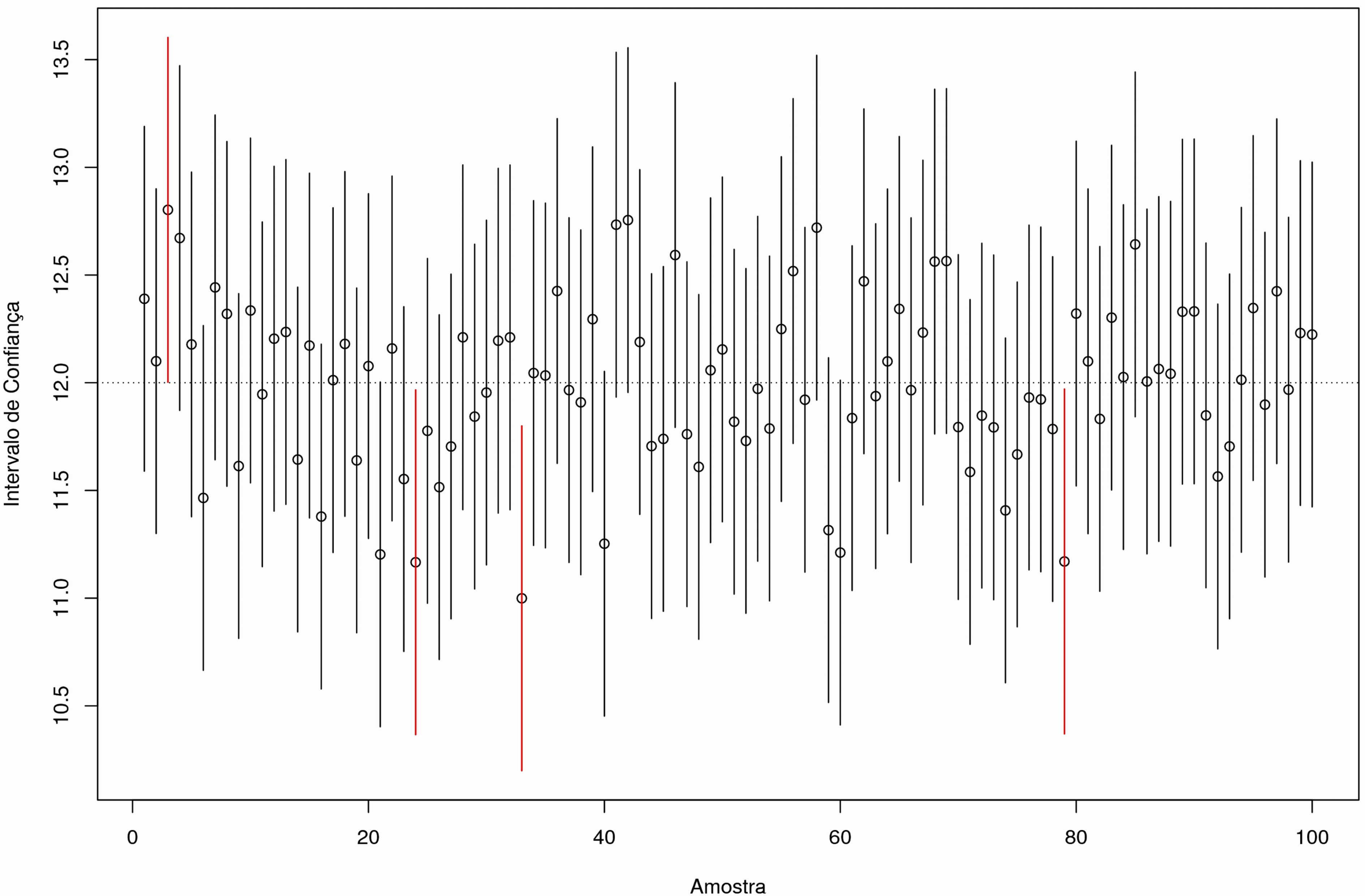
$$P\left(-1{,}96\times0{,}41\leqslant\bar{X}-12\leqslant1{,}96\times0{,}41\right)=0{,}95$$

$$P\left(-\bar{X}-1{,}96\times0{,}41\leqslant-12\leqslant-\bar{X}+1{,}96\times0{,}41\right)=0{,}95$$

$$P\left(\bar{X}+1{,}96\times0{,}41\geqslant12\geqslant\bar{X}-1{,}96\times0{,}41\right)=0{,}95$$

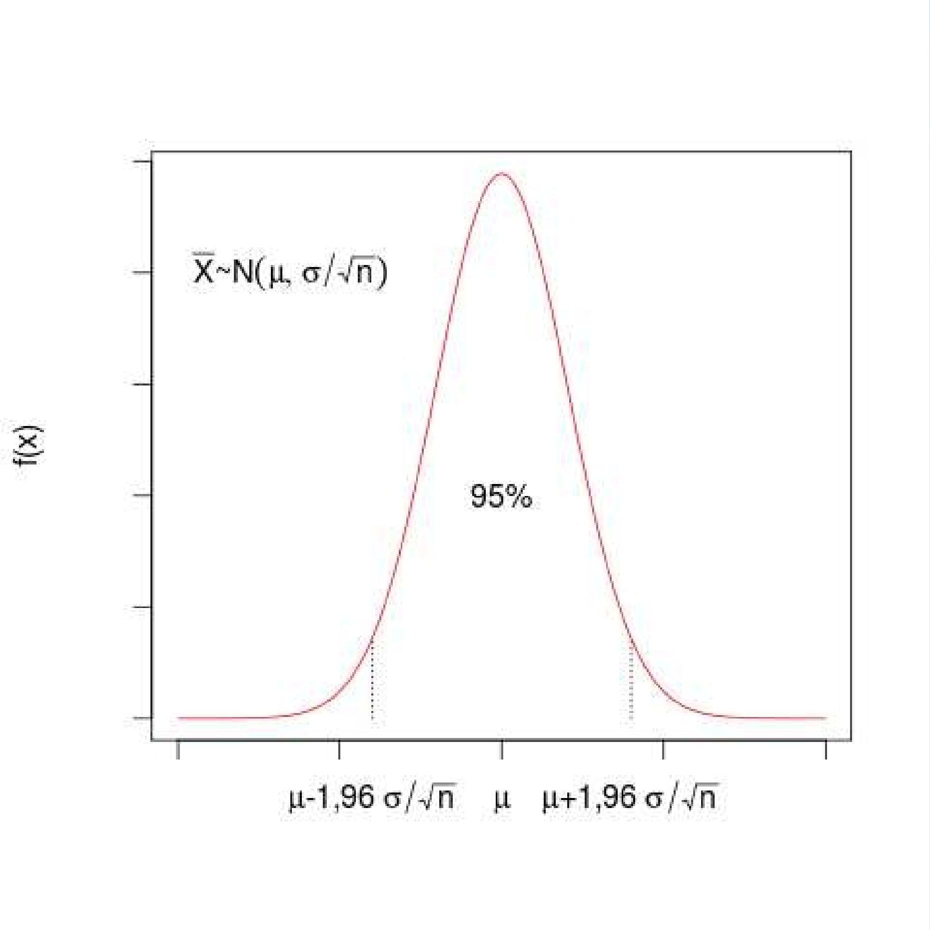
$$P\left(\bar{X}-1{,}96\times0{,}41\leqslant12\leqslant\bar{X}+1{,}96\times0{,}41\right)=0{,}95$$

- 95% dos intervalos ( $\bar{X}$ ±1,96x0,41) cobrem a média populacional 12



Intervalo de Confiança
Amostra
13.5
13.0
12.5
12.0
11.5
11.0
10.5
0
20
40
60
80
100

# Generalizando




- 95% das médias amostrais estão entre (μ±1,96 σ/√n)

$$P(\mu - 1{,}96\,\sigma/\sqrt{n} \leqslant \bar{X} \leqslant \mu + 1{,}96\,\sigma/\sqrt{n}) = 0{,}95$$

$$\Downarrow$$

$$P(\bar{X} - 1{,}96\,\sigma/\sqrt{n} \leqslant \mu \leqslant \bar{X} + 1{,}96\,\sigma/\sqrt{n}) = 0{,}95$$

- 95% dos intervalos ( $\bar{X}$ ±1,96 σ/√n) cobrem μ

# Intervalo de confiança para μ

- Quando σ é conhecido usando TCL podemos construir um intervalo para estimar a média populacional μ

- IC de 95% para a média populacional μ

$$\left( \bar{X} - 1{,}96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1{,}96 \frac{\sigma}{\sqrt{n}} \right)$$

# Exemplo

- Intervalo de confiança de 95% para taxa de hemoglobina média assumindo σ conhecido e igual a 1

# t-Student

- Na prática σ também não é conhecido!!!

- Então σ é estimado usando s

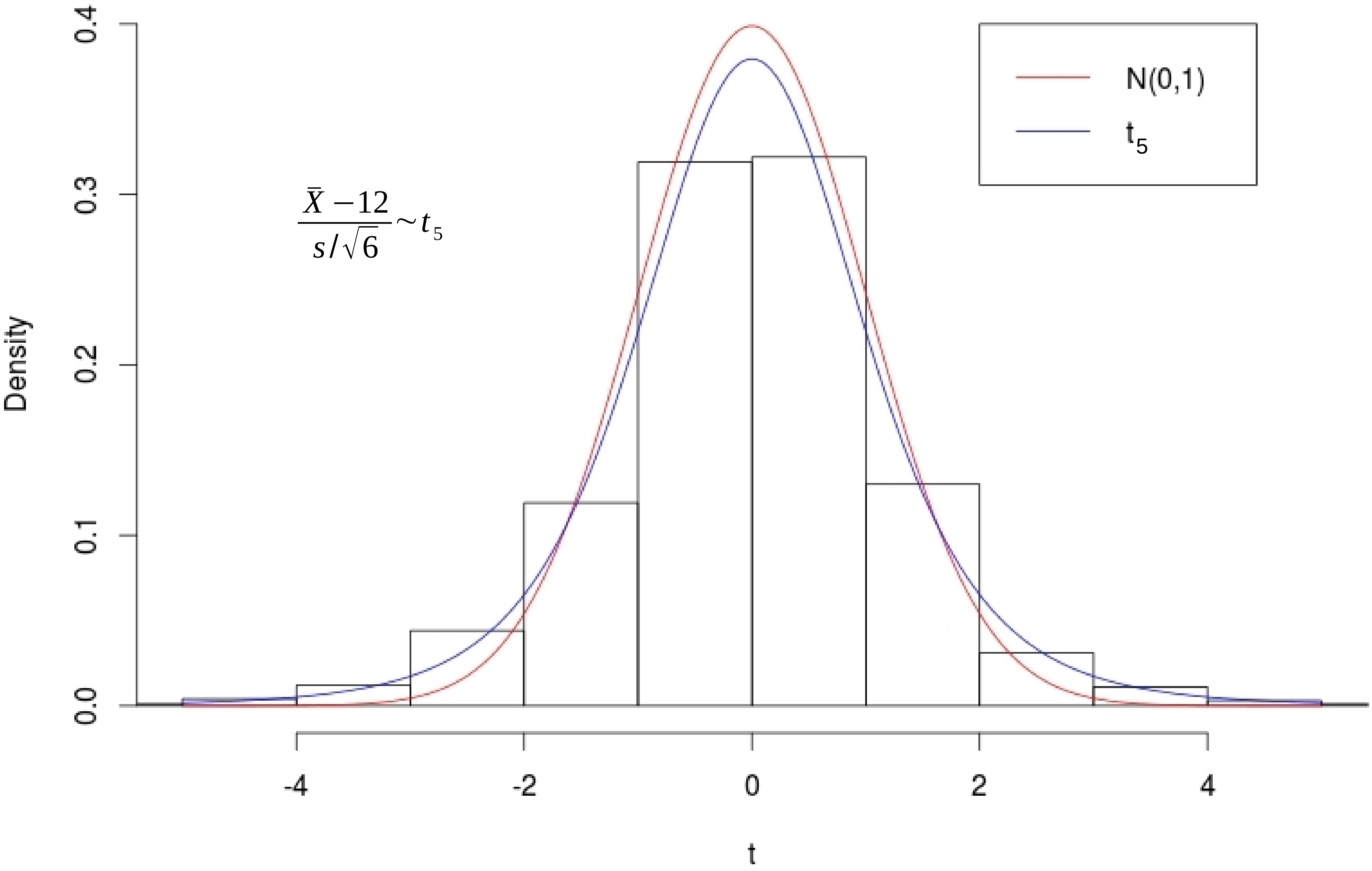
$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim N(0,1)$$

# t-Student

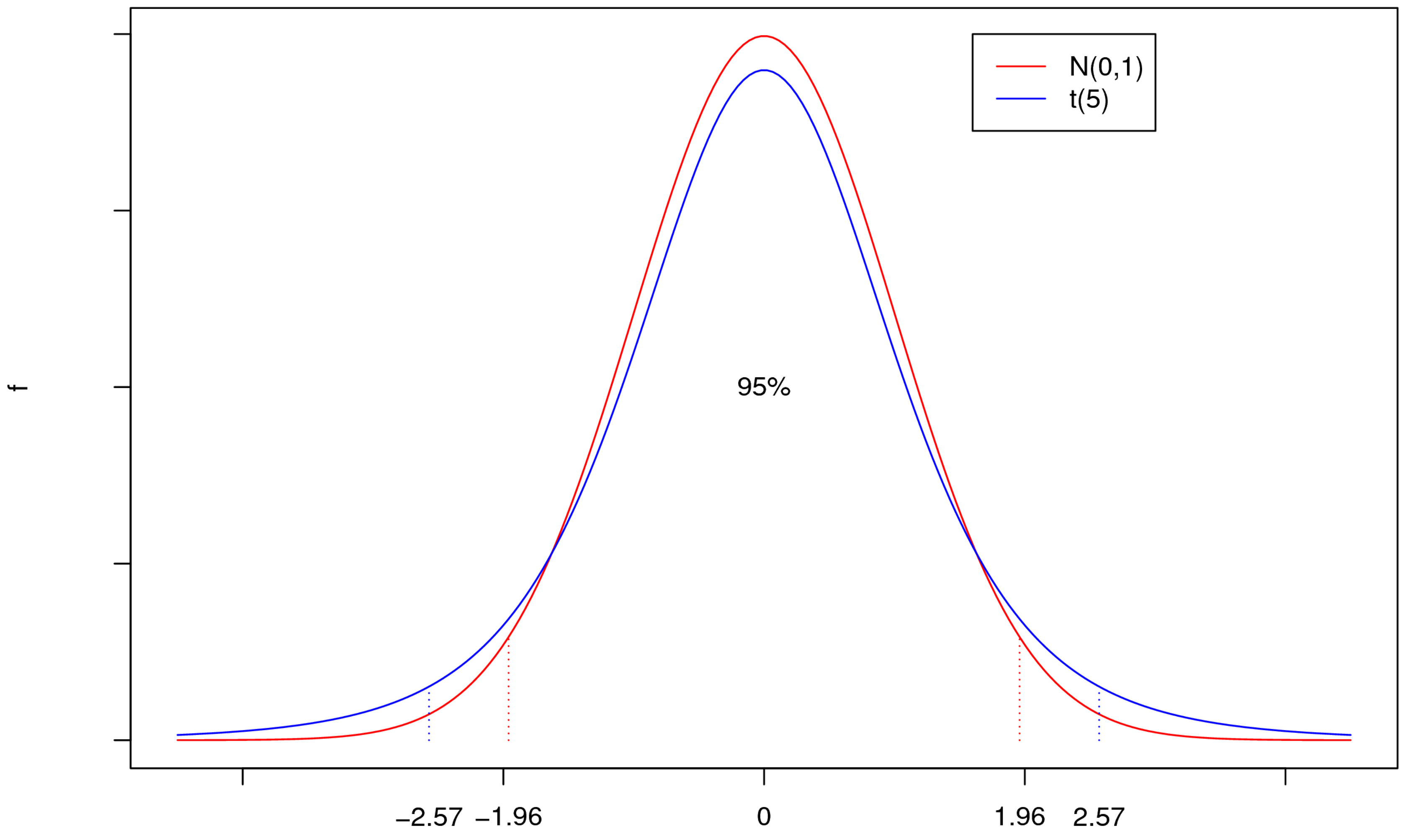
- Na prática σ também não é conhecido!!!

- Então σ é estimado usando s

$$\cancel{\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim N(0,1)} \Rightarrow \boxed{\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}}$$



$\frac{\bar{X}-12}{s/\sqrt{6}} \sim t_5$
N(0,1)
$t_5$
Density
t
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4



N(0,1)
t(5)
95%
f
−2.57
−1.96
0
1.96
2.57

# t-Student

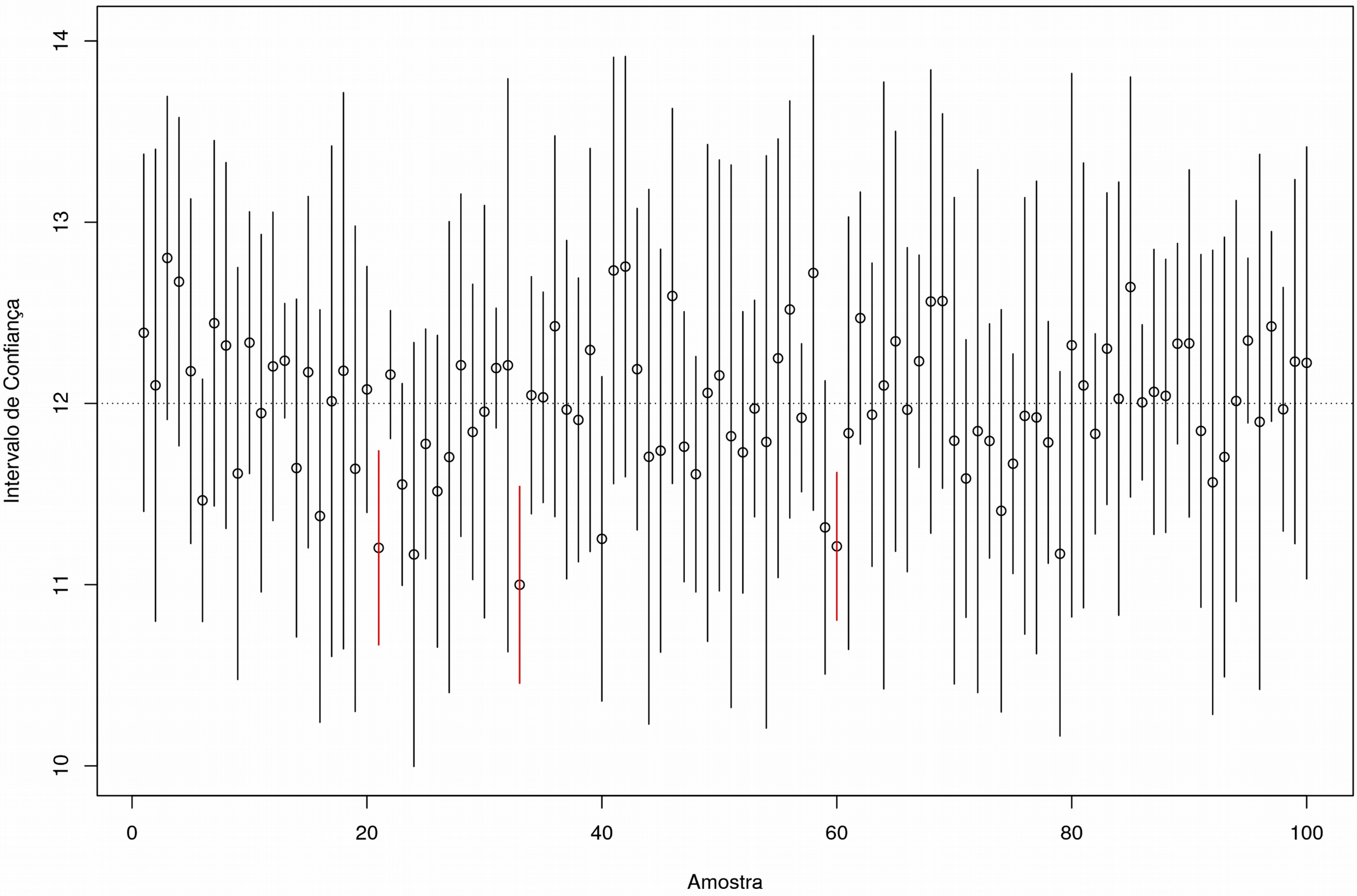
- Na prática σ também não é conhecido!!!

- Então σ é estimado usando s

$$\cancel{\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim N(0,1)} \Rightarrow \boxed{\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}}$$

- IC para a média populacional μ

$$\boxed{\left(\bar{X}-t_{n-1}\frac{s}{\sqrt{n}}, \bar{X}+t_{n-1}\frac{s}{\sqrt{n}}\right)}$$



Intervalo de Confiança
Amostra
10
11
12
13
14
0
20
40
60
80
100

# Exemplo

- Intervalo de confiança de 95% para a taxa de hemoglobina média estimando $\sigma$ usando s

# Intervalo de confiança para uma proporção

- Devido ao Teorema Central do Limite, para n grande e p não muito próximo de 0 ou 1, a distribuição da proporção amostral $\hat{p}$ será aproximadamente normal com média p e um desvio-padrão

$$EP = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Este resultado pode ser usado para construir um intervalo de confiança para a verdadeira proporção p.
- O intervalo de confiança de aproximadamente 95% para p é

$$\boxed{\hat{p} \pm 1{,}96 \times EP}$$

- Uma regra geral é que este intervalo de confiança é válido quando tanto $n\hat{p}$ quanto $n(1-\hat{p})$ forem maiores do que 10.

# Exemplo

- Um ensaio clínico foi realizado para determinar a preferência entre dois analgésicos, A e B, contra dor de cabeça. Cem pacientes que sofrem de dor de cabeça crônica receberam em dois tempos diferentes o analgésico A e o analgésico B.
- A ordem na qual os pacientes receberam os analgésicos foi determinada ao acaso. Os pacientes desconheciam esta ordem.
- Ao final do estudo foi perguntado a cada paciente qual analgésico lhe proporcionou maior alívio: o primeiro ou o segundo. Dos 100 pacientes, 45 preferiram A e 55 preferiram B.
- Baseado nestas informações podemos dizer que há prefência por algum dos analgésicos?

# Exemplo

- Dizemos que não há preferência por um dos analgésicos quando a proporção dos que preferem A ($p_A$), é igual a proporção dos que preferem B ($p_B$). Como temos dois resultados possíveis, $p_A$ e $p_B$ são iguais quando $p_A=p_B=0,5$.
- Um intervalo de 95% de confiança para a verdadeira proporção de pacientes que preferem o analgésico A é:

$$\left(0,45 \pm 1,96\sqrt{0,45 \times 0,55/100}\right) = \left(0,35\,;\,0,55\right)$$

- Então com 95% de confiança, a verdadeira proporção de pacientes que preferem o analgésico A está entre 0,35 e 0,55. Observe que este intervalo contem o valor 0,5 então concluímos que não existem evidências amostrais de preferência por um dos analgésicos.

# Dimensionamento de amostras

- Sabemos construir intervalos para alguns parâmetros populacionais (média e proporção)
- Em ambos os casos, fixamos o nível de confiança de acordo com a probabilidade de acerto que desejamos ter na estimação por intervalo.
- O nível de confiança pode ser aumentado até tão próximo de 100% quanto se queira, mas isso resultará em intervalos de amplitude cada vez maiores, o que significa perda de precisão na estimação.
- Seria desejável intervalos com alto nível de confiança e grande precisão. Isso porém requer uma amostra suficientemente grande, pois, para n fixo, a confiança e a precisão variam em sentidos opostos.
- Veremos a seguir como determinar o tamanho das amostras nos casos de estimação da média ou de uma proporção populacional.

# Dimensionamento de amostras

- Vimos que o intervalo de confiança de 95% para a média μ da população quando σ é conhecido tem semi-amplitude (ou precisão) d dada pela expressão

$$d = 1{,}96 \times \frac{\sigma}{\sqrt{n}}$$

- O problema resolvido foi:
  - Fixados o nível de confiança de 95% e n, determinar d.
- É evidente dessa expressão que podemos resolver outro problema:
  - Fixados, d (ou seja, fixada a precisão) e o nível de confiança, determinar n.

$$n=\left(\frac{1{,}96\times\sigma}{d}\right)^2$$

- Não conhecendo o desvio-padrão da população, deveriamos substituí-lo pelo desvio-padrão amostral s e usar t de Student ao invés de 1,96.
- Porém não tendo ainda sido retirada a amostra, não dispomos do valor de s. Se não conhecemos nem ao menos um limite superior para σ, a única solução será colher uma amostra-piloto de $n_0$ elementos para, com base nela obtermos uma estimativa de s, empregando a seguir a expressão:

$$n=\left(\frac{t_{n_0-1}\times s}{d}\right)^2$$

- Se $n \leq n_0$, a amostra-piloto já terá sido suficiente para a estimação. Caso contrário, deveremos retirar, ainda, da população os elementos necessários à complementação do tamanho mínimo de amostra.

- Procedemos de forma análoga se desejamos estimar uma proporção populacional com determinada confiança e dada precisão. No caso de população suposta infinita, da expressão

$$d = 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

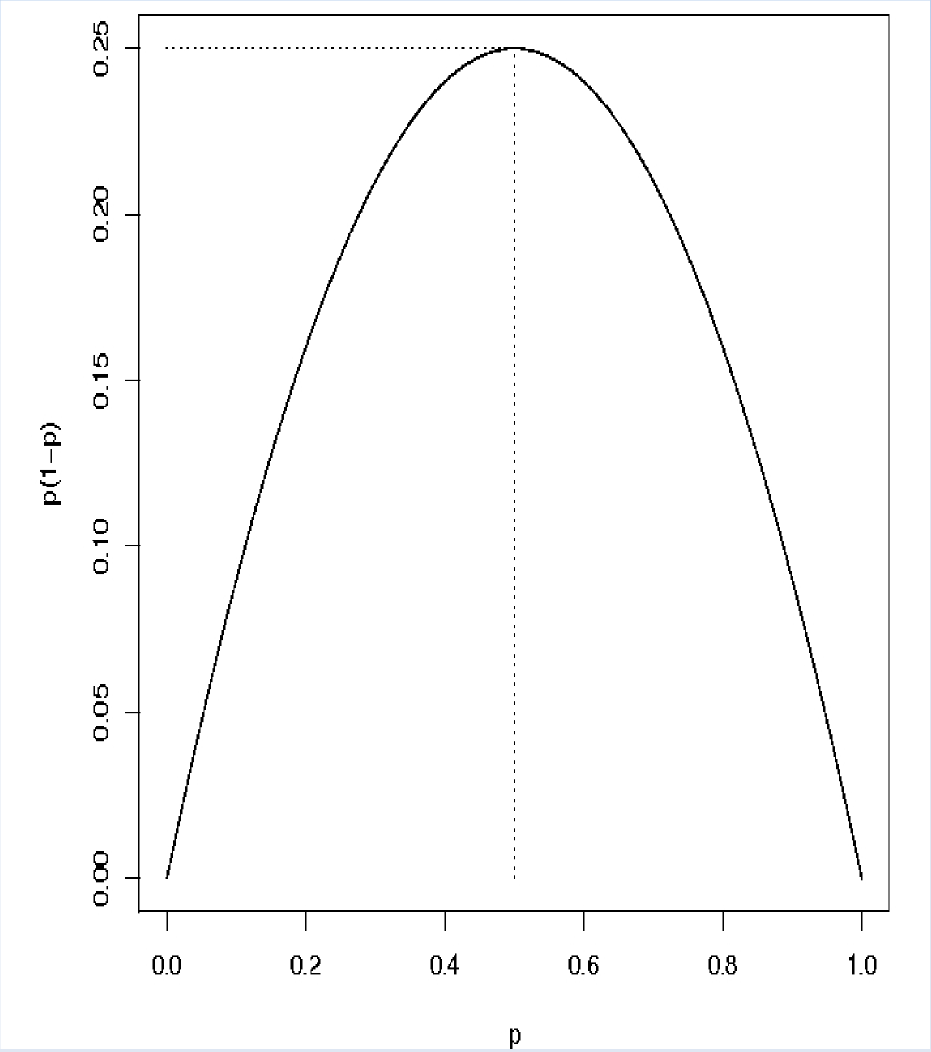
podemos obter

$$\boxed{n = \left(\frac{1{,}96}{d}\right)^2 \hat{p}(1-\hat{p})}$$

- A determinação do tamanho de amostra depende de valores desconhecidos de p.

- Essa dificuldade pode ser resolvida através de uma amostra-piloto, ou analisando-se o comportamento do fator p(1-p).

- Vê-se da figura que p(1-p) é a expressão de uma parábola cujo ponto de máximo é p=0,5.



- Se substituirmos, p(1-p) por seu valor máximo, 1/4, seguramente o tamanho de amostra obtido será suficiente para a estimação de qualquer que seja p. Isso equivale a considerar

$$n=\left(\frac{1{,}96}{d}\right)^2\frac{1}{4}=\left(\frac{1{,}96}{2d}\right)^2$$

- Evidentemente, usando-se essa expressão corre-se o risco de se superdimensionar a amostra. Isso ocorrerá se p for na realidade próximo de 0 ou 1. Se o custo envolvido for elevado e proporcional ao tamanho de amostra, é mais prudente a tomada de uma amostra-piloto.

# Exercícios

- Qual o tamanho de amostra necessário para se estimar a média de uma população infinita cujo desvio-padrão é igual a 4, com 95% de confiança e precisão de 0,5?
- Qual o tamanho de amostra suficiente para estimarmos a proporção de pessoas doentes que precisam de tratamento, com precisão de 0,02 e 95% de confiança, sabendo que essa proporção seguramente não é superior a 0,2?