CE008
Introdução à Bioestatística

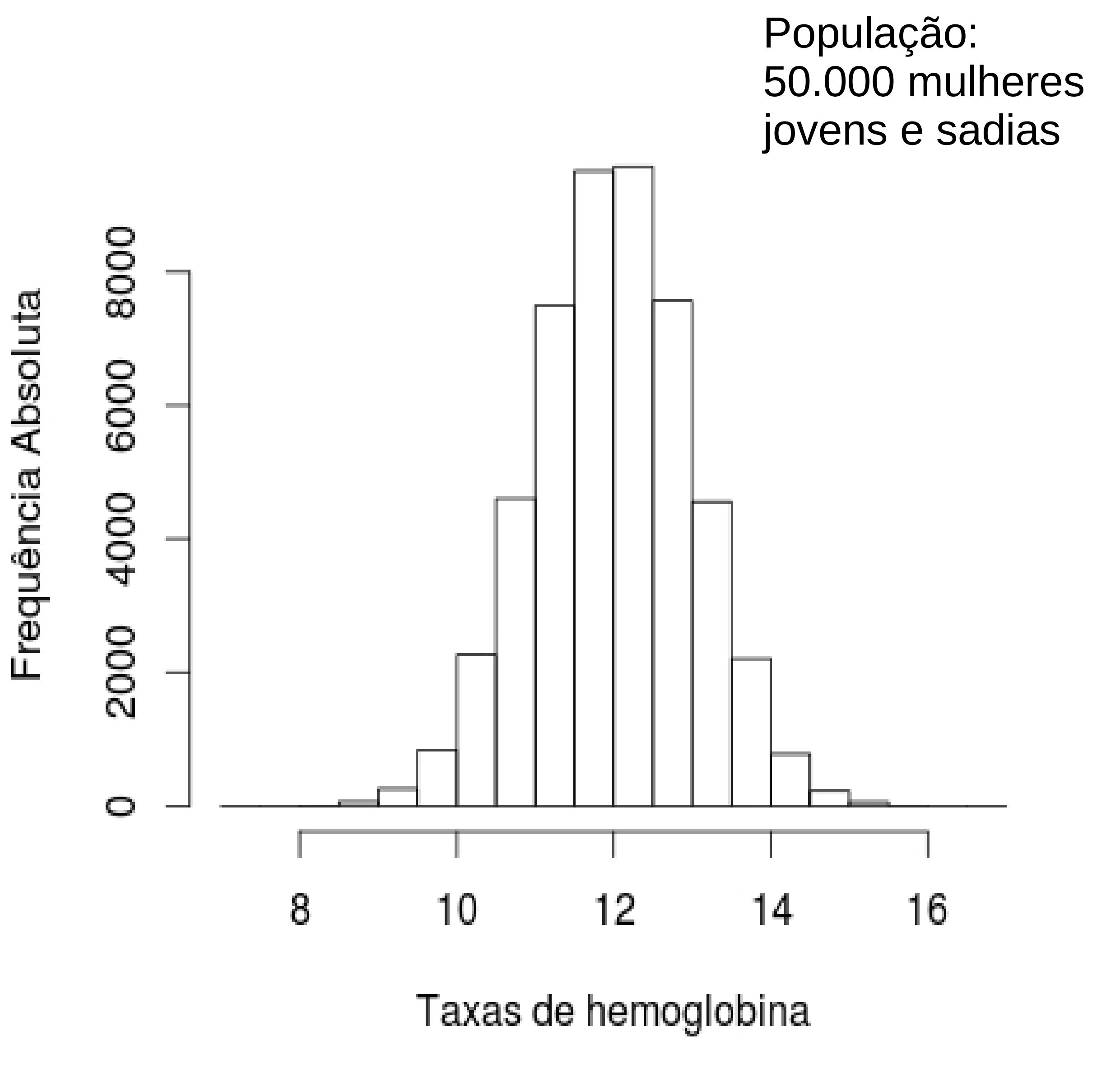
INFERÊNCIA ESTATÍSTICA

Silvia Shimakura

Estimação

- Amostras são usadas para estimar quantidades desconhecidas de uma população.
- Exemplo: prevalência de doenças, efeito de medicamentos, diferença entre grupos
- É importante saber qual é a variação destas estimativas de amostra para amostra.
- Teoria de probabilidades permite usar amostras para estimar quantidades de populações, e determinar a precisão destas estimativas.

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.

Estimação de médias

- O que acontece quando retiramos uma amostra de uma população e estimamos a média populacional usando a média amostral?
- A estimativa é precisa?
- E se retiramos outra amostra a estimativa da média populacional será diferente?

Amostragem 1

- Uma amostra de tamanho 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?

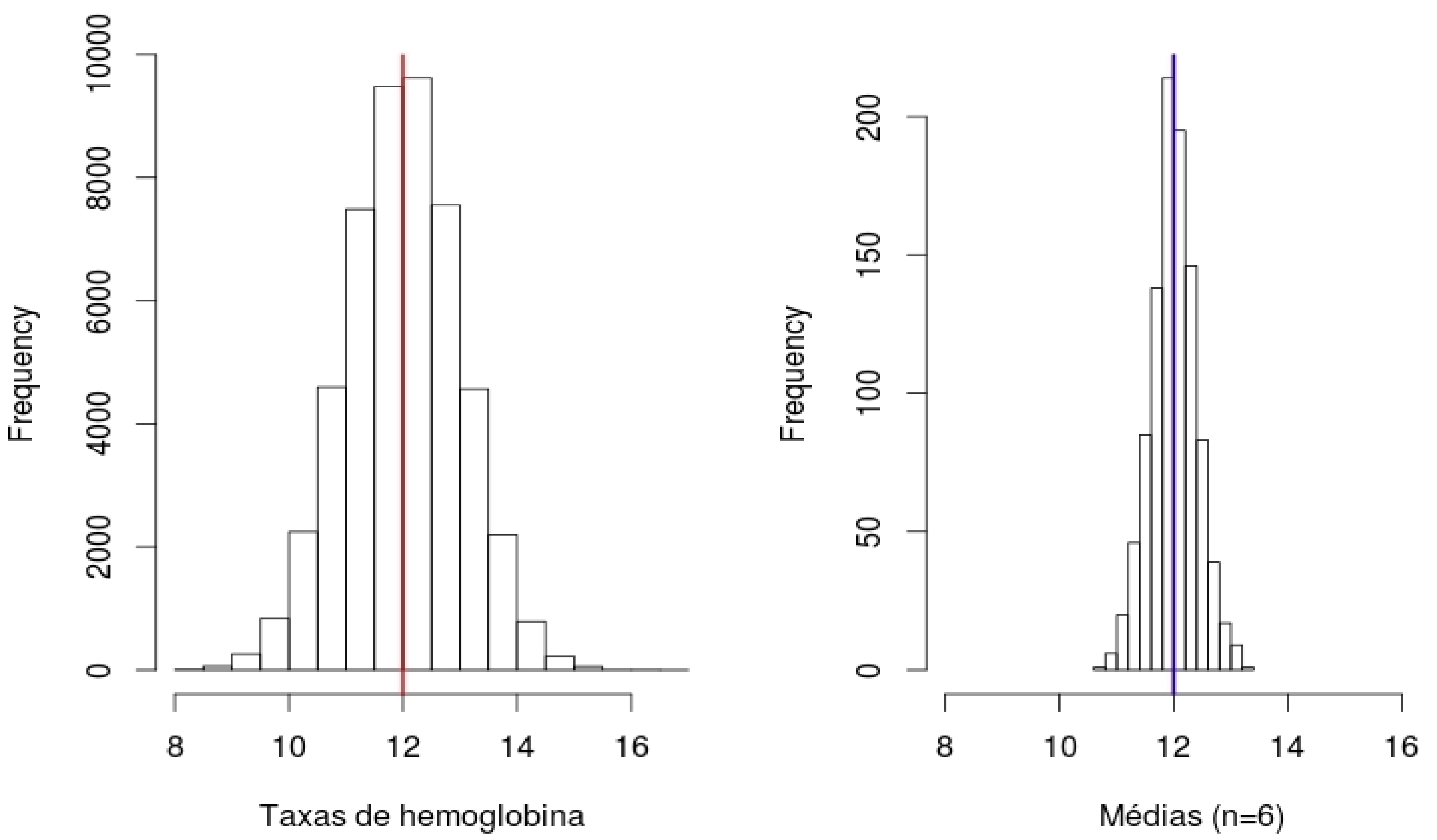
RESPOSTA

- 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



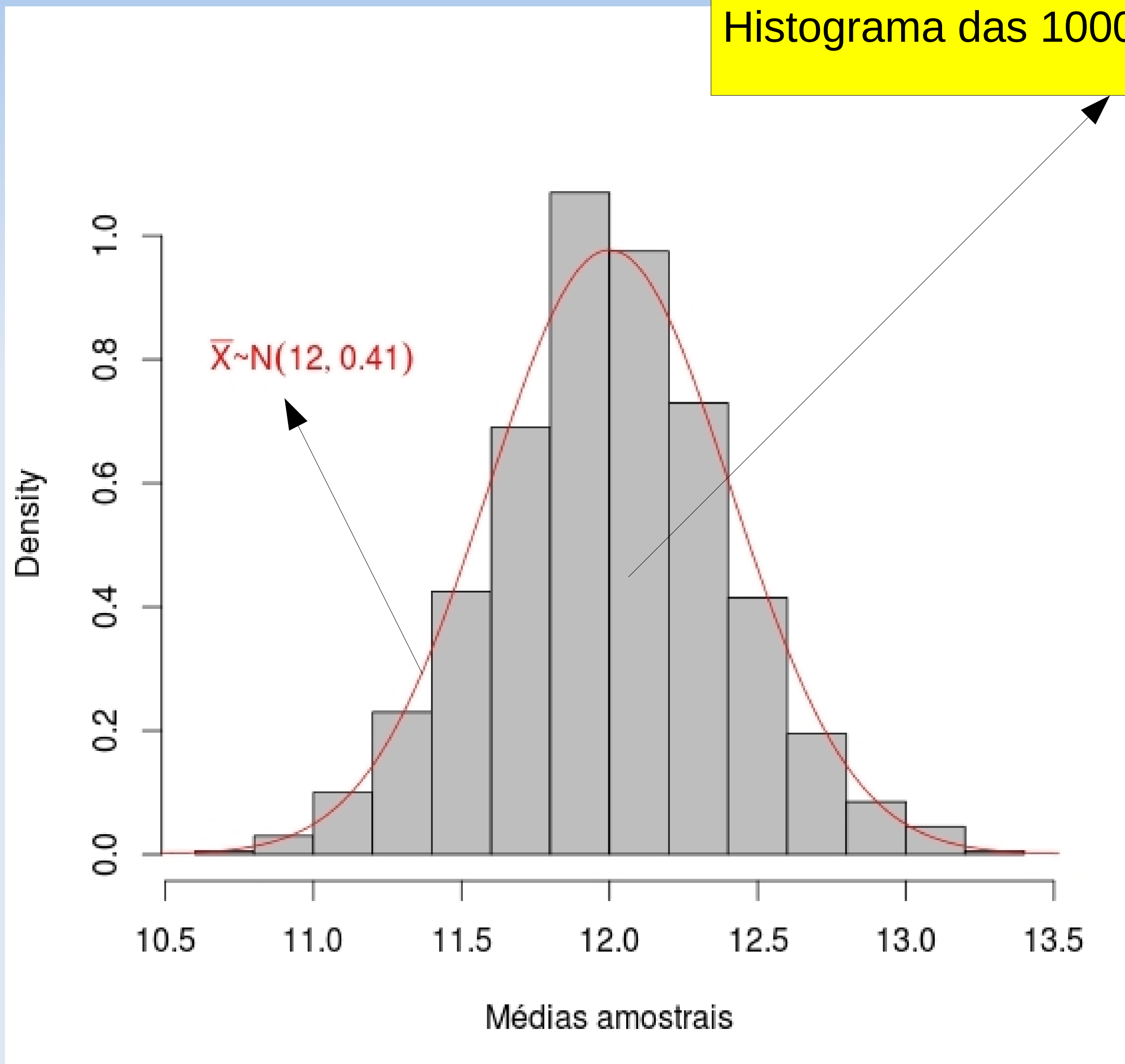
Erro padrão da média amostral

- As 1000 médias podem ser usadas para estimar os parâmetros da distribuição de $\bar{X}$
- Média das 1000 médias amostrais = 12
- Desvio-padrão das 1000 médias amostrais = 0,41 < 1
- Teorema Central do Limite: a distribuição das médias amostrais é Normal com média igual à média da população e desvio-padrão

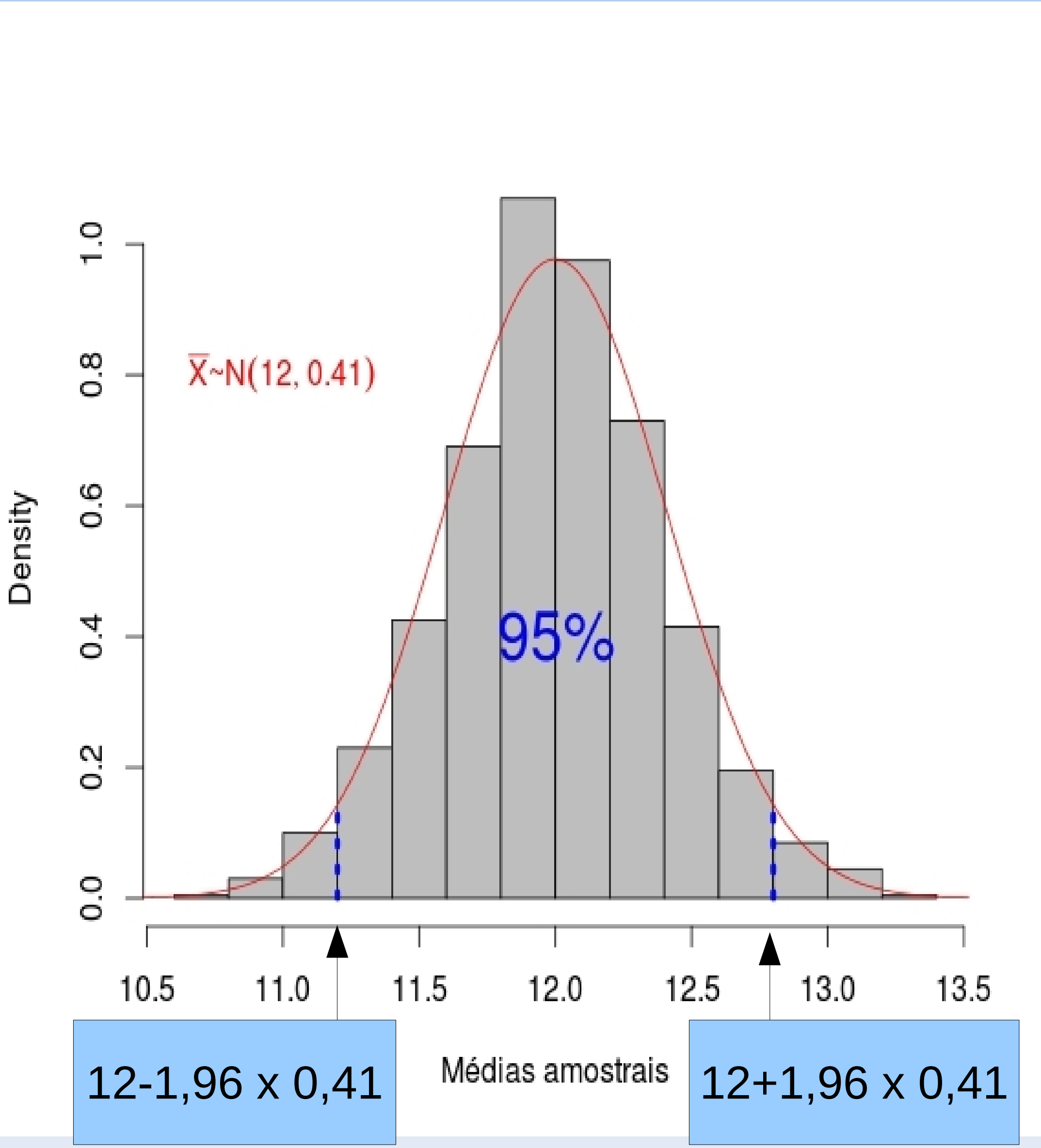
$$\sigma / \sqrt{n} = 1 / \sqrt{6} = 0{,}41$$

Teorema Central do Limite

Histograma das 1000 médias amostrais



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 a média populacional 12

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

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

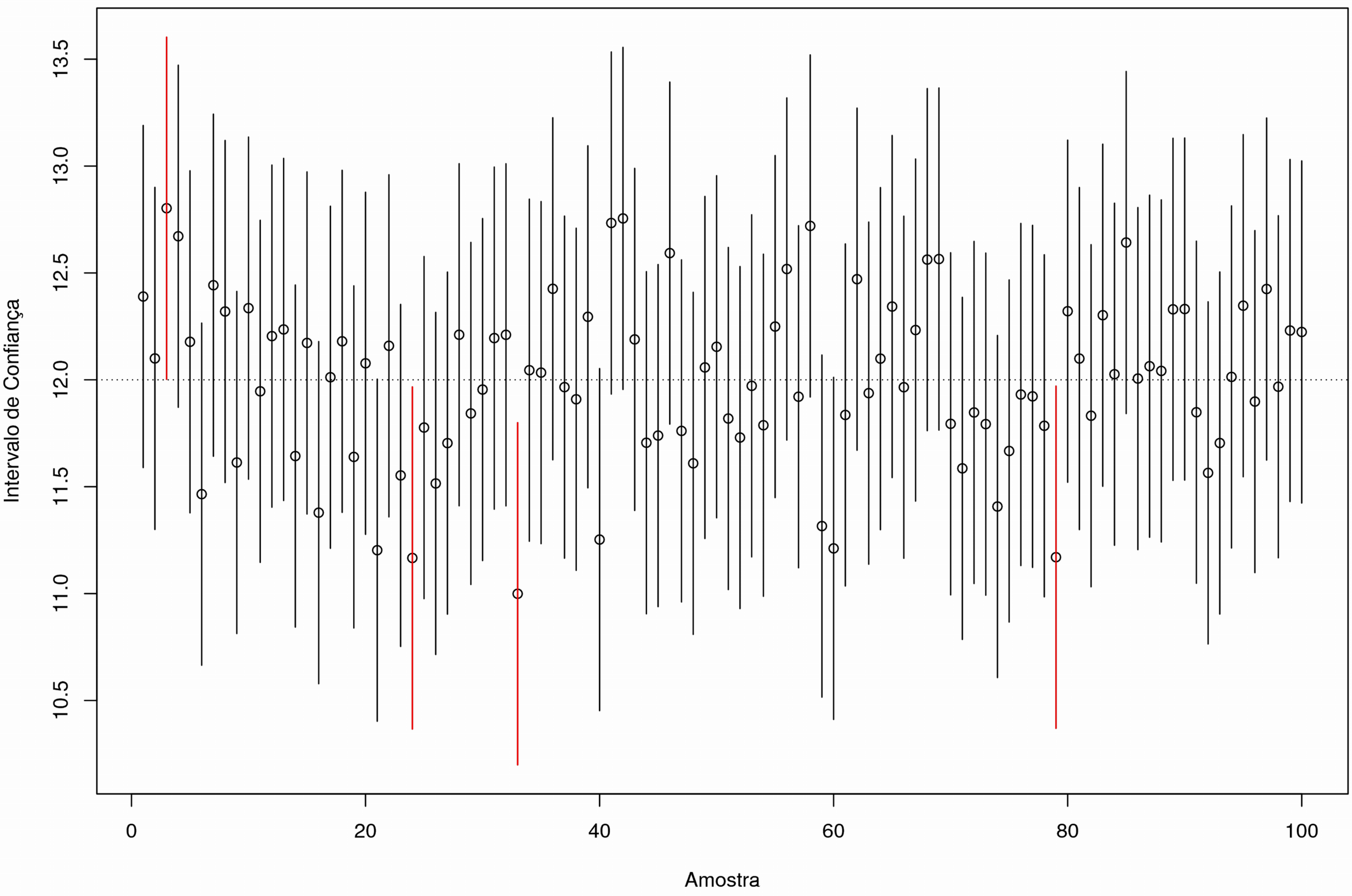
$$P\left(-1{,}96\times 0{,}41\leqslant \bar{X}-12\leqslant 1{,}96\times 0{,}41\right)=0{,}95$$

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

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

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

- 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

Teorema central do limite

- Usando o TCL, podemos obter uma estimativa intervalar para a média populacional µ

- Intervalo de confiança 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

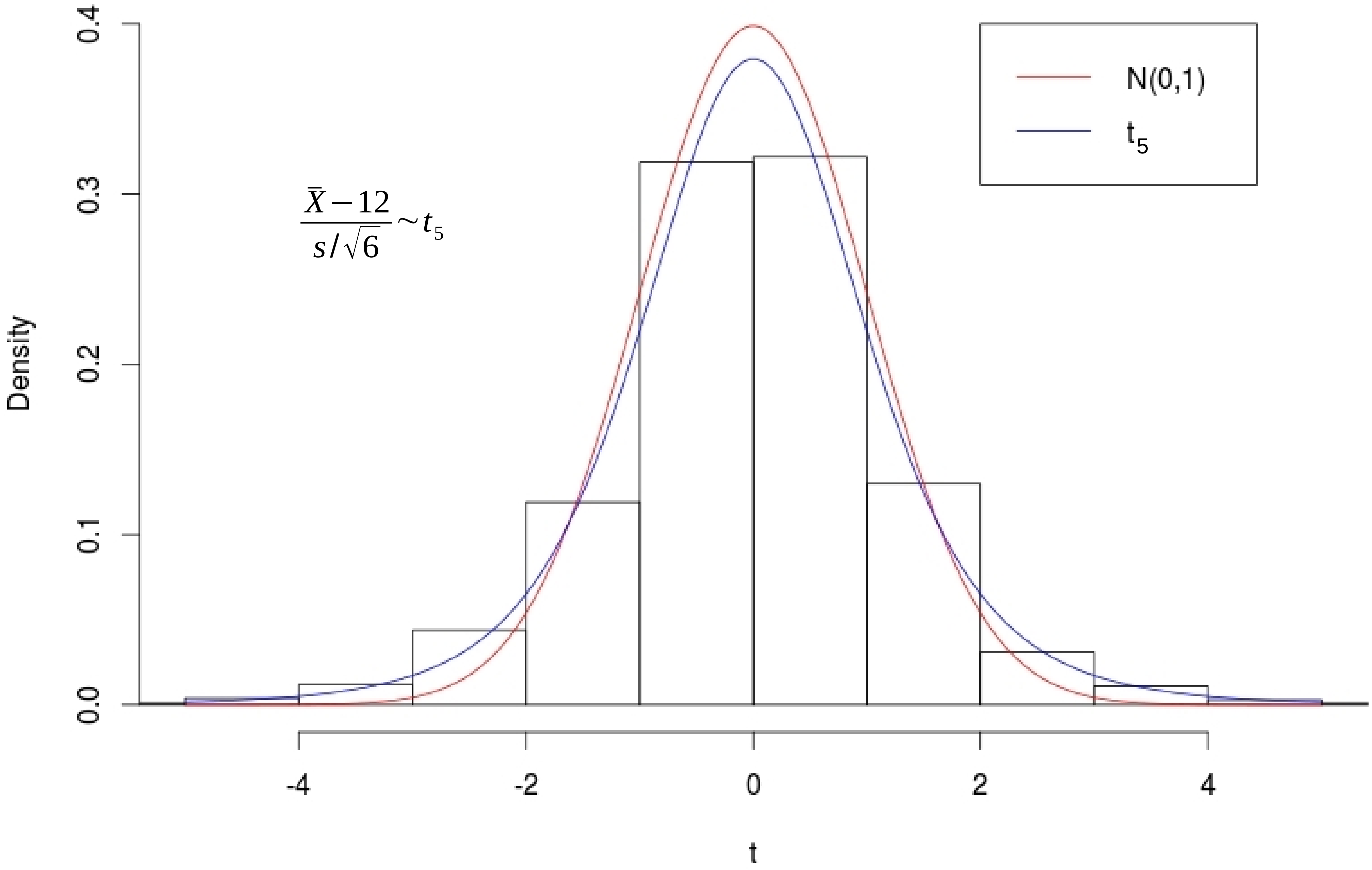
t-Student

- Pelo TCL $\bar{X} \sim N\left(\mu, \sigma/\sqrt{n}\right)$ então:

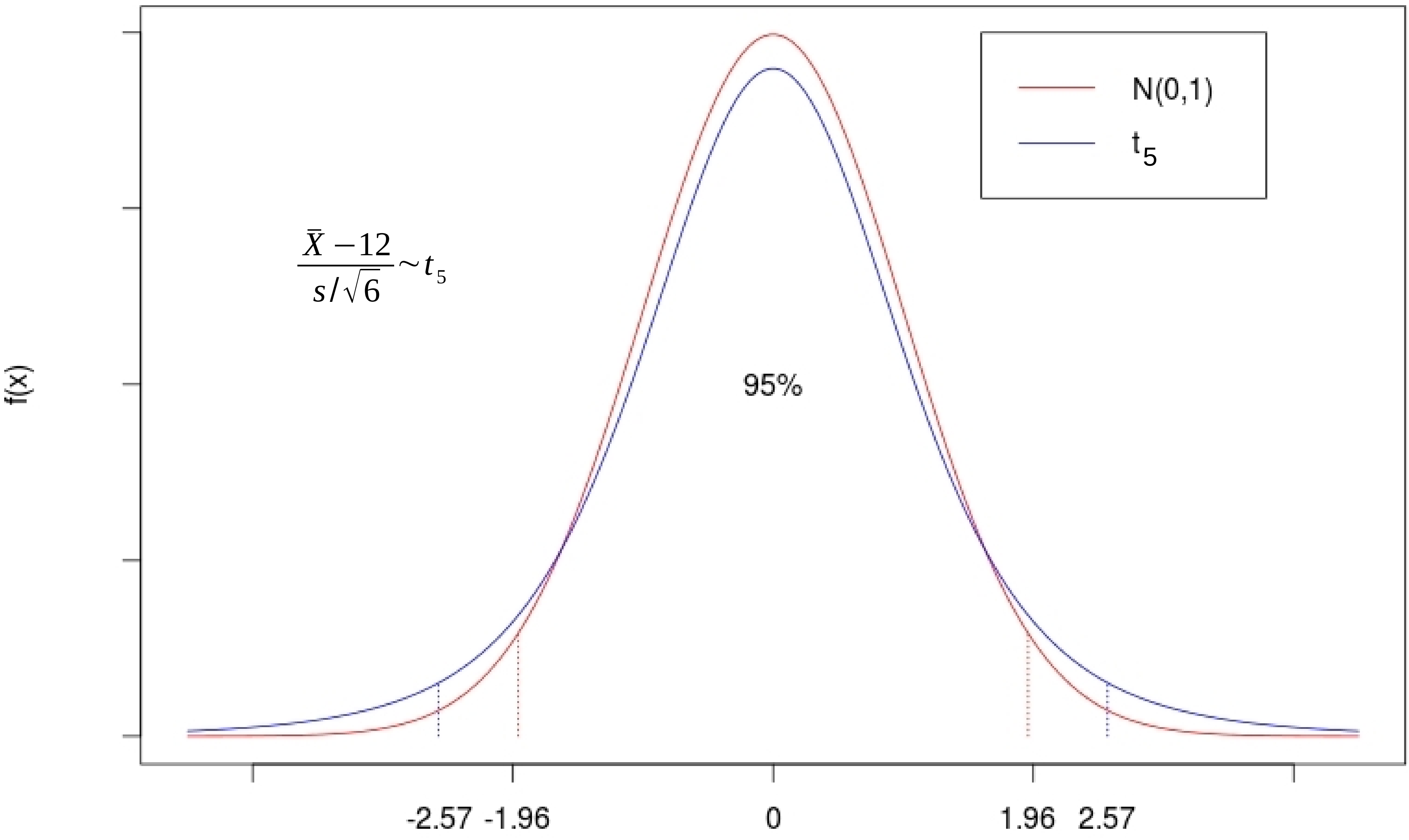
$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

- Na prática σ não é conhecido então se estimarmos σ :

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t_{n-1}$$



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

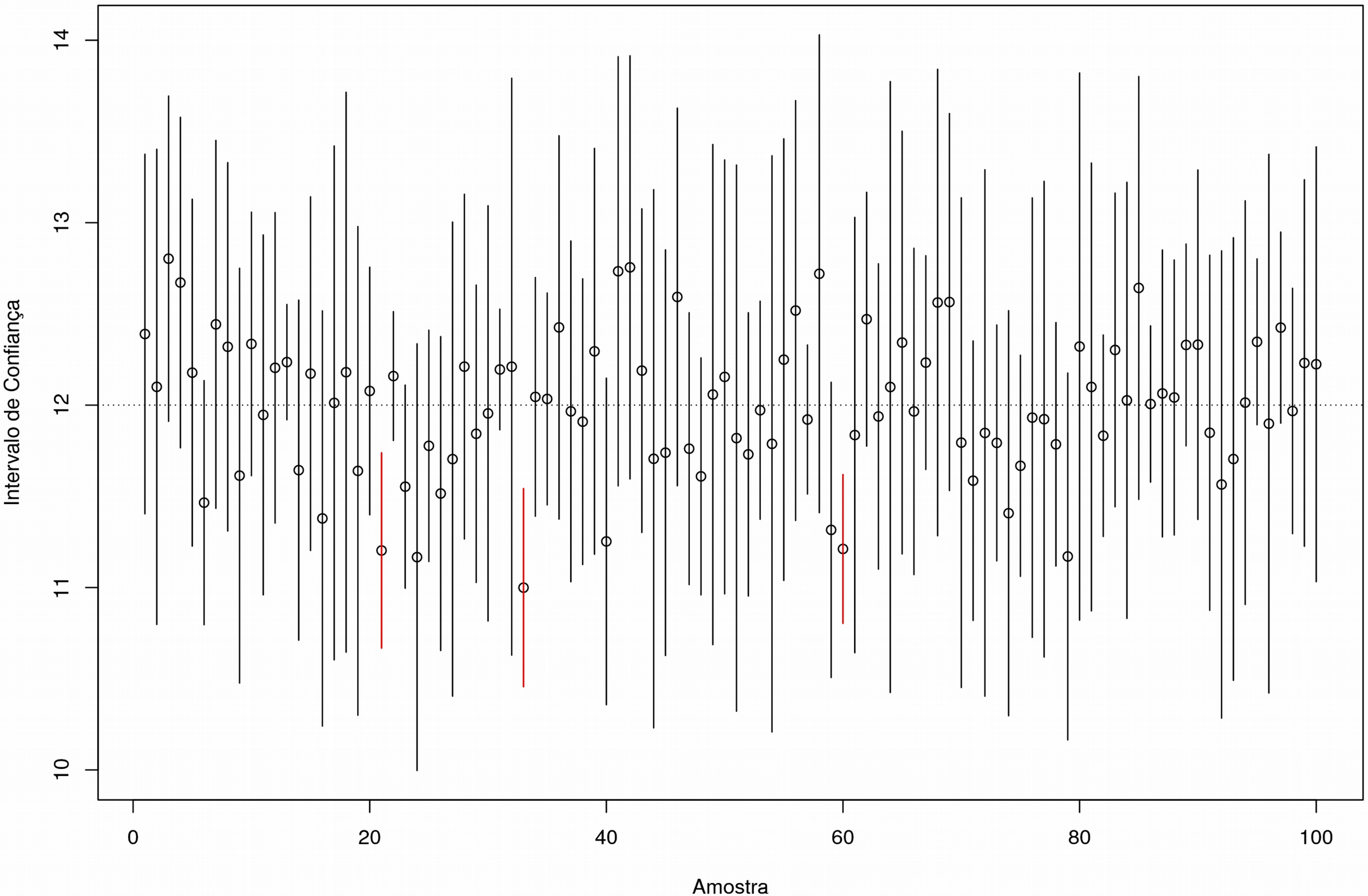


N(0,1)
t_5
$\frac{\bar{X}-12}{s/\sqrt{6}} \sim t_5$
95%
f(x)
-2.57 -1.96
0
1.96 2.57

t-Student

- Intervalo de confiança de 95% para a média populacional μ quando σ é desconhecido

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



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 σ usando s

Intervalo de confiança para uma proporção

- Os resultados obtidos para a média amostral podem ser estendidos para a proporção amostral $\hat{p}$
- Usando o TCL, para n grande, a distribuição da proporção amostral $\hat{p}$ será proximadamente normal com média p e um desvio-padrão

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

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

$$\boxed{\left[\hat{p}\pm 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]}$$

- Uma regra geral é que este intervalo de confiança é válido quando $n\hat{p}>10$ e $n(1-\hat{p})>10$

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. 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 estatísticas 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 cobertura desejada para 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 implica em perda de precisão na estimação.
- Seria desejável um intervalo 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 o nível de confiança e a precisão d, determinar n.

Dimensionamento de amostras

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

- 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, uma solução é colher uma amostra-piloto de n_0 elementos e com base nela obtermos uma estimativa de σ.
- O tamanho de amostra será então calculado usando a expressão:

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

Dimensionamento de amostras

- 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

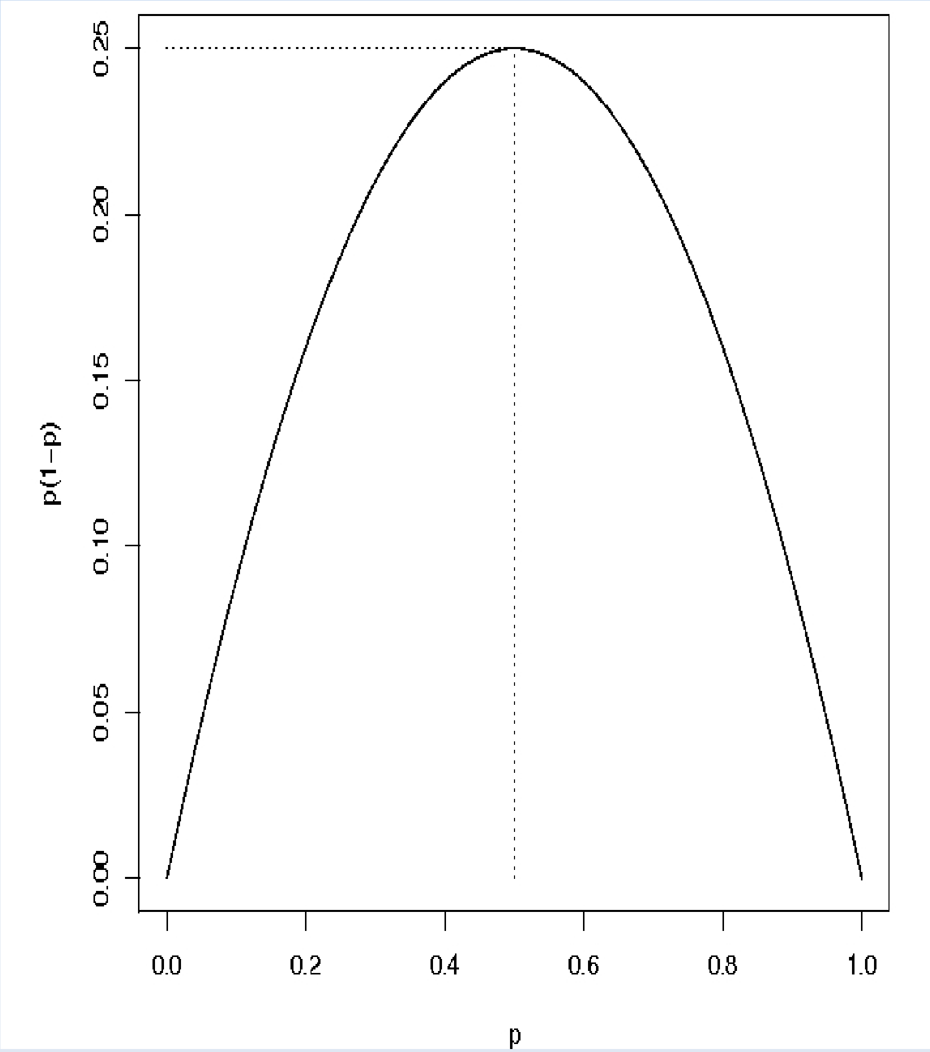
$$d = 1{,}96\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

podemos obter

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

- A determinação do tamanho de amostra depende de uma estimativa 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 a função p(1-p) é a expressão de uma parábola cujo ponto de máximo é p=0,5.



Dimensionamento de amostras

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

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

- Evidentemente, há o risco de 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?