Estimativas

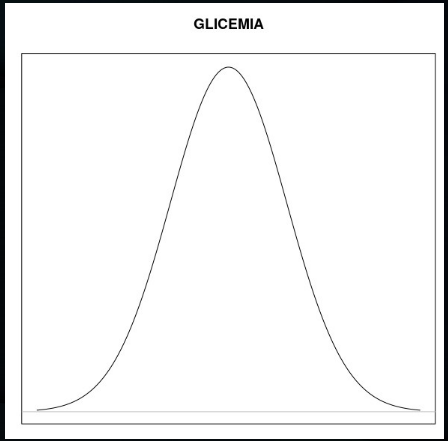
- Como estimar a média de glicemia (de animais saudáveis) de uma população?
- Como estimar a proporção de pacientes que morrem de uma dada doença?

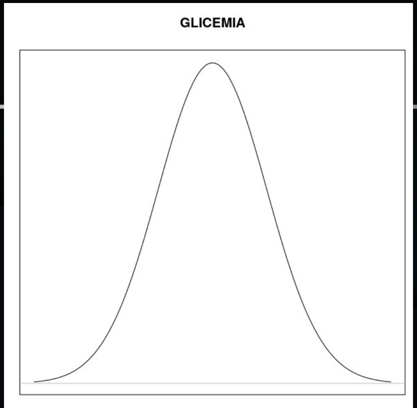
Algumas notações

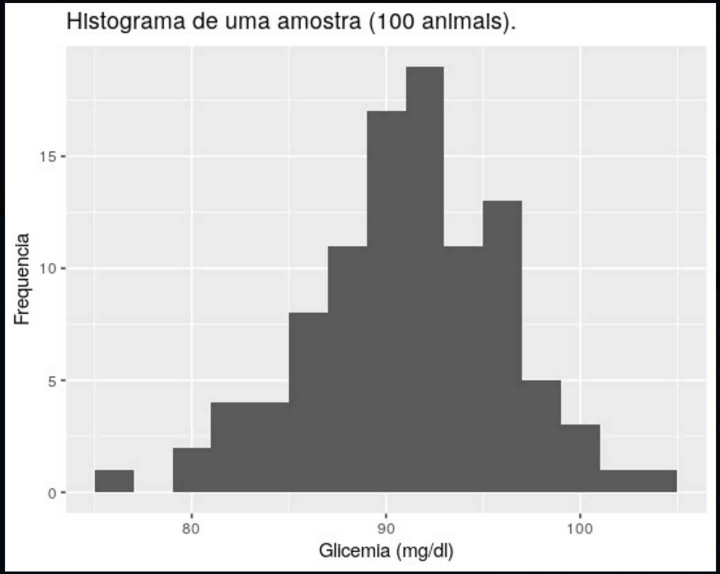
Parâmetro	População	Amostra
Média	μ	$\bar{x}$
Desvio-padrão	σ	s
Variância	σ^2	s^2
Proporção	p	$\hat{p}$
Tamanho	N	n

Variáveis quantitativas
Estimativa de média

- População tem média μ desconhecida
- Como, a partir de uma amostra, é possível estimar μ ?



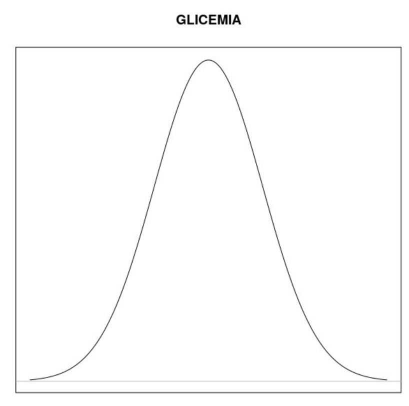




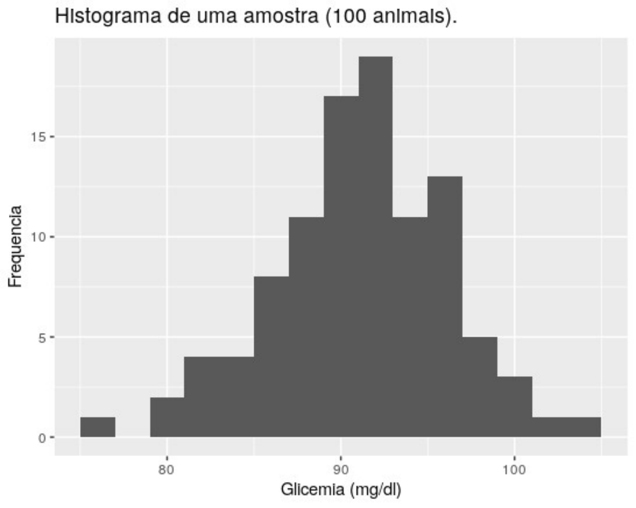
Exemplo
População: μ = ?

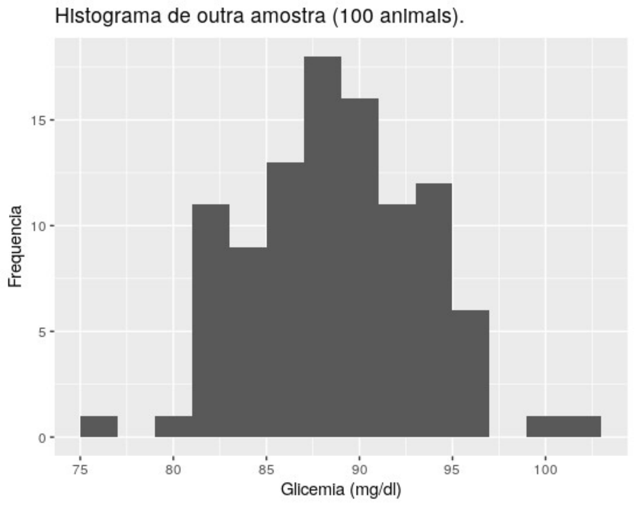
Amostra: $\bar{x}$ = 91,08

Animal	Glicemia
1	95,00
2	88,50
3	99,70
...	...
100	85,40



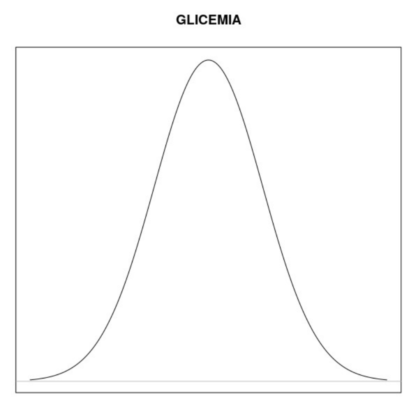
População: μ = ?



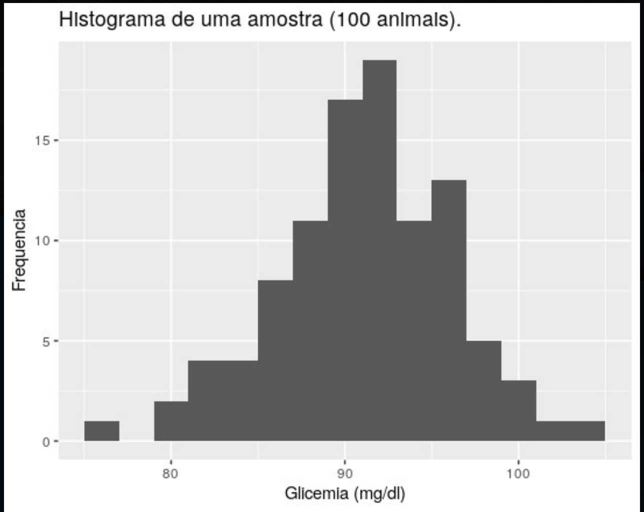


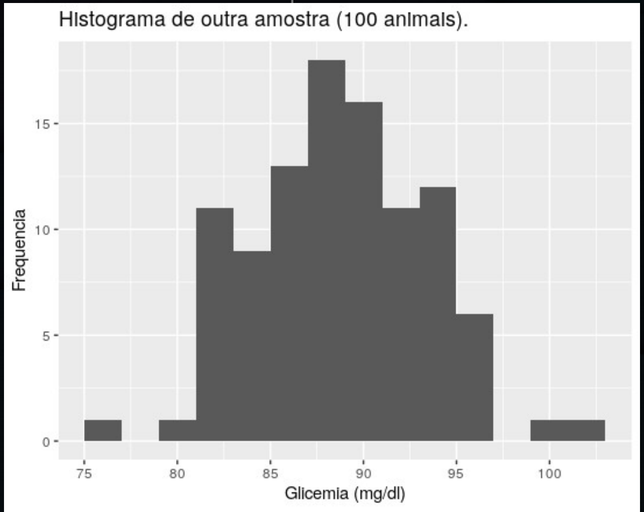
Amostra 1: $\bar{x}$ = 91,08

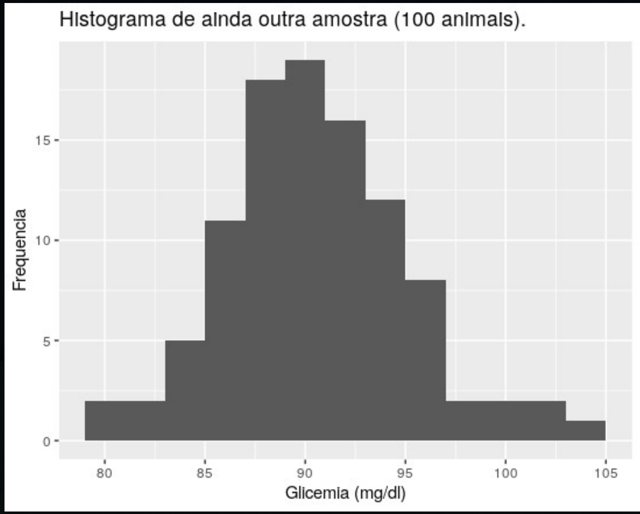
Amostra 2: $\bar{x}$ = 88,79



População: μ = ?





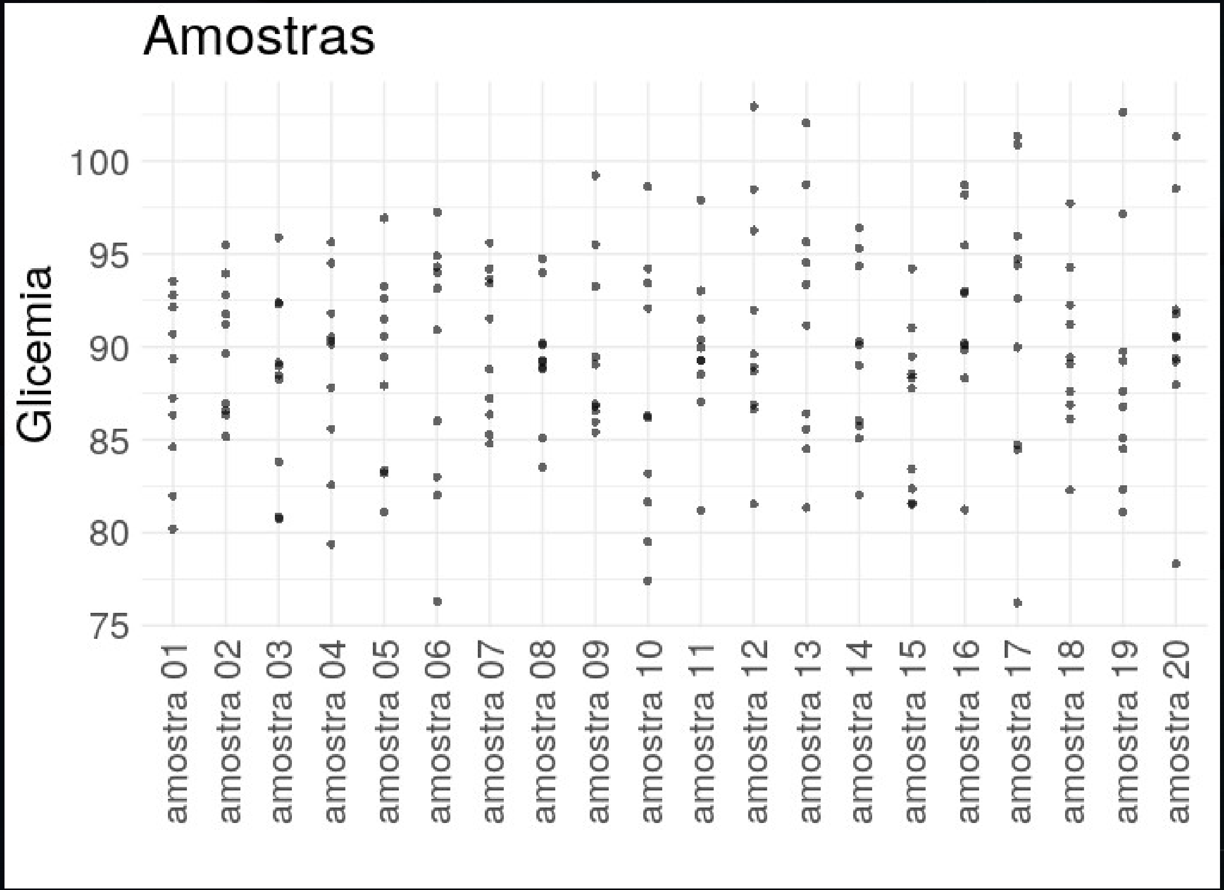


Amostra 1: $\overline{x}$ = 91,08

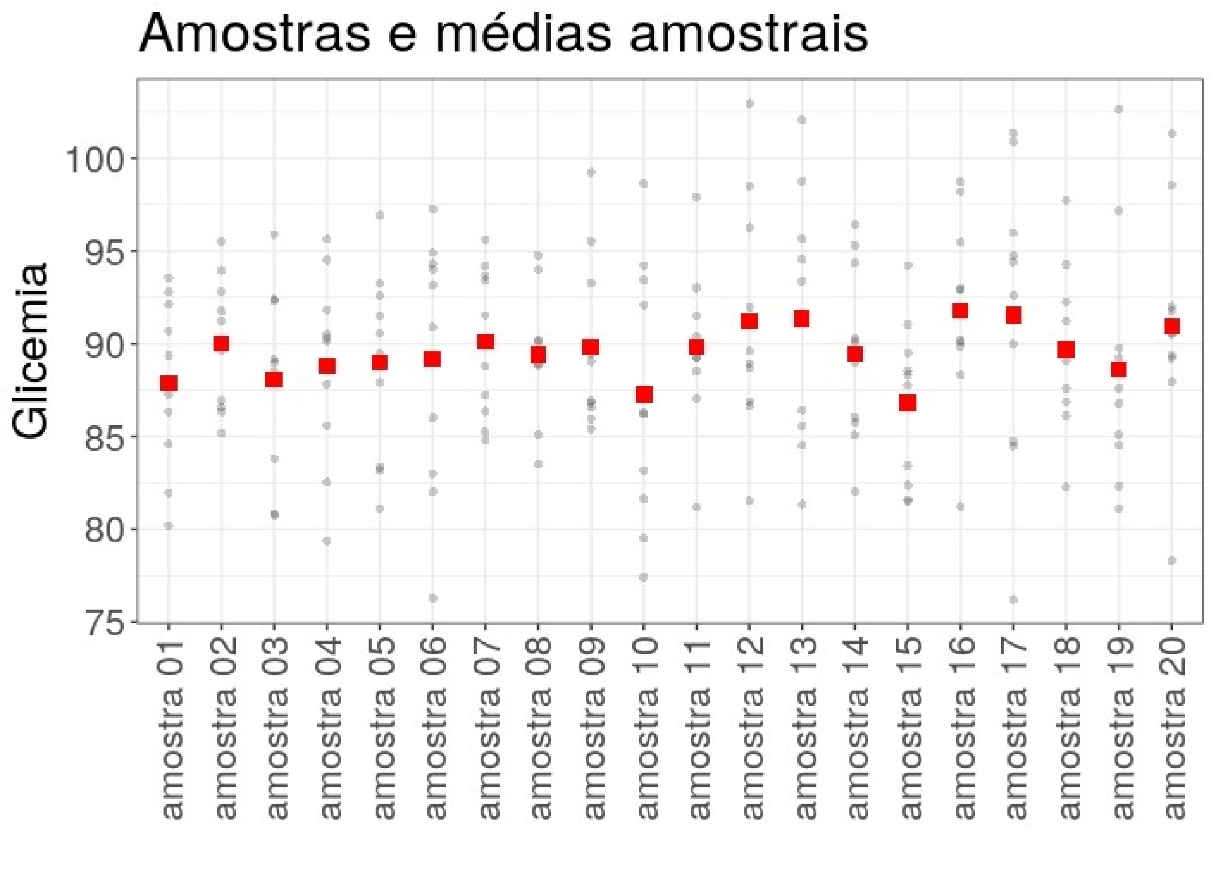
Amostra 2: $\overline{x}$ = 88,79

Amostra 3: $\overline{x}$ = 90,61

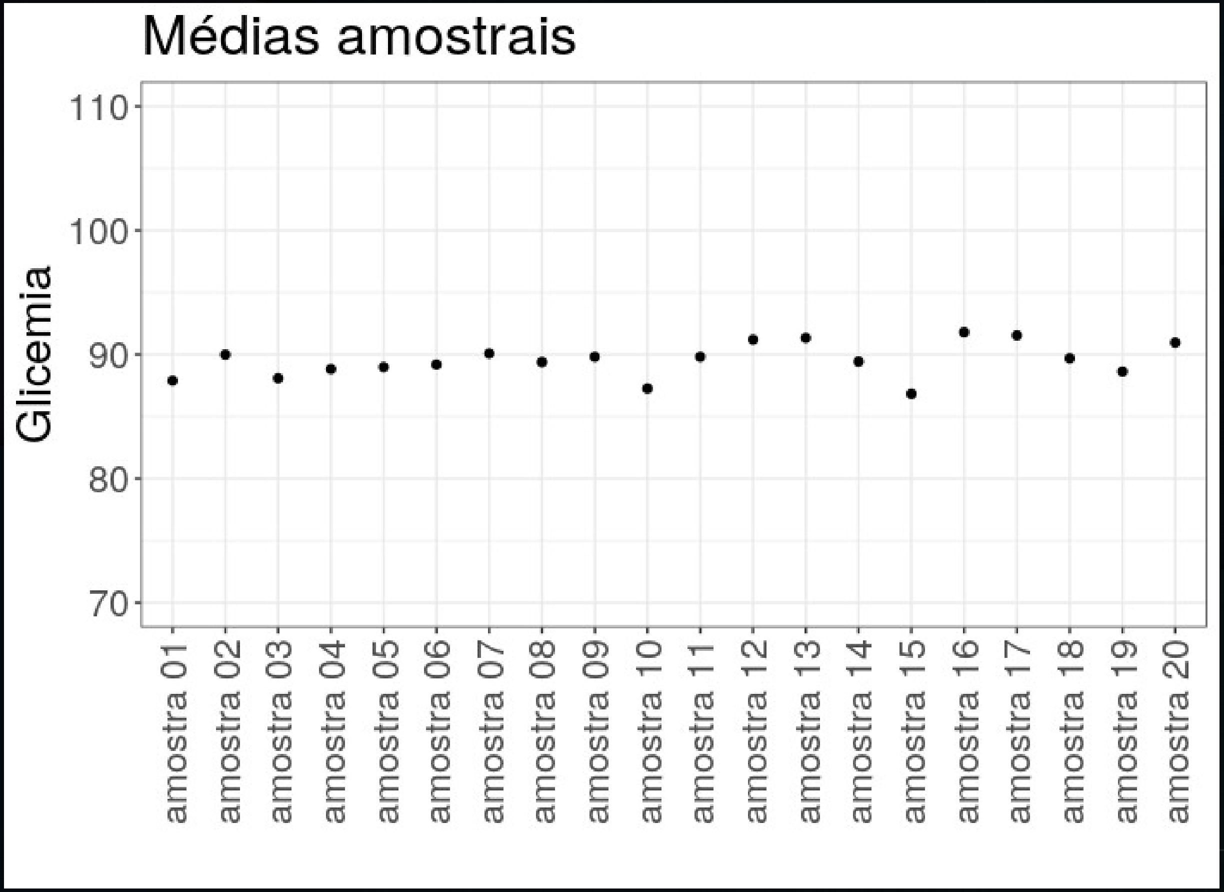
Com 20 amostras



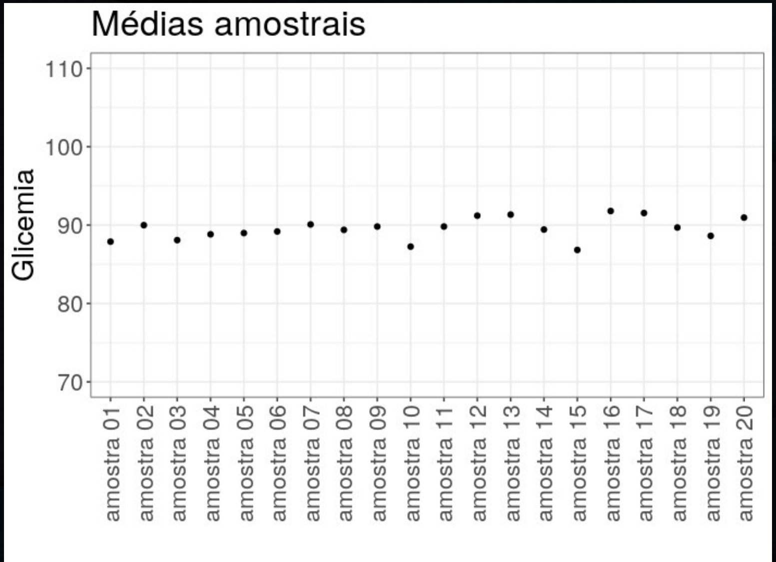
Com 20 amostras

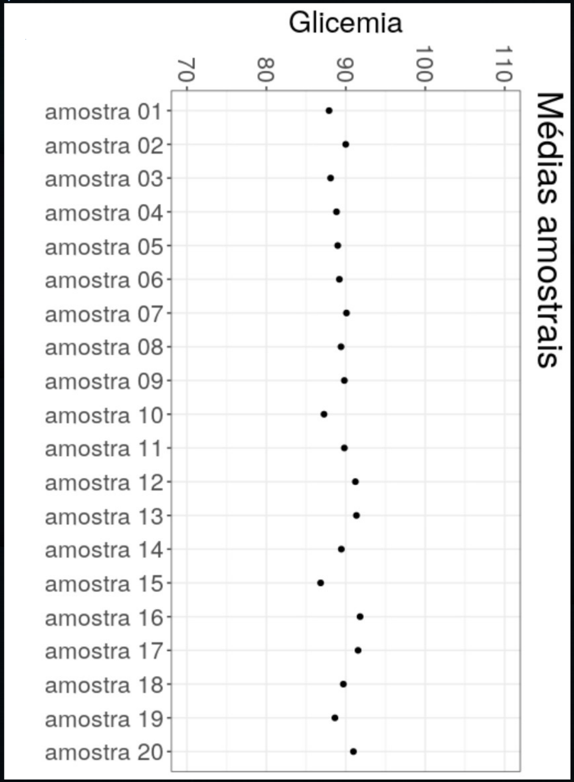


Com 20 amostras

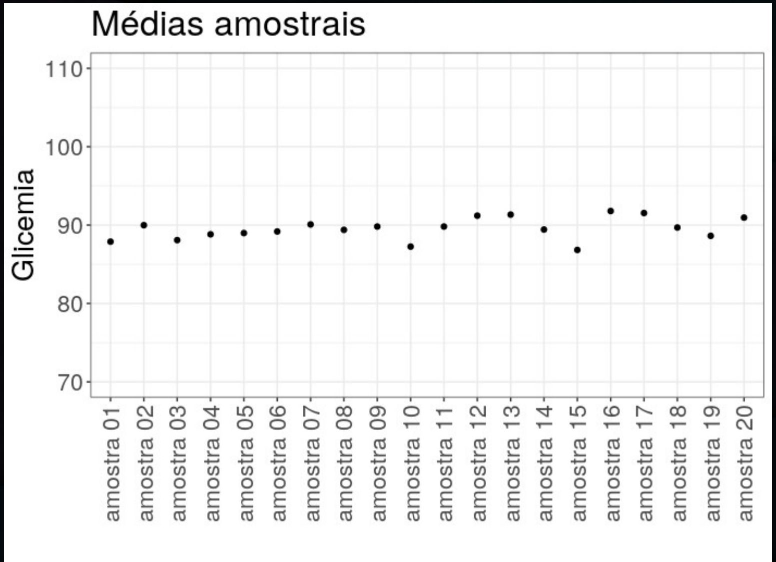


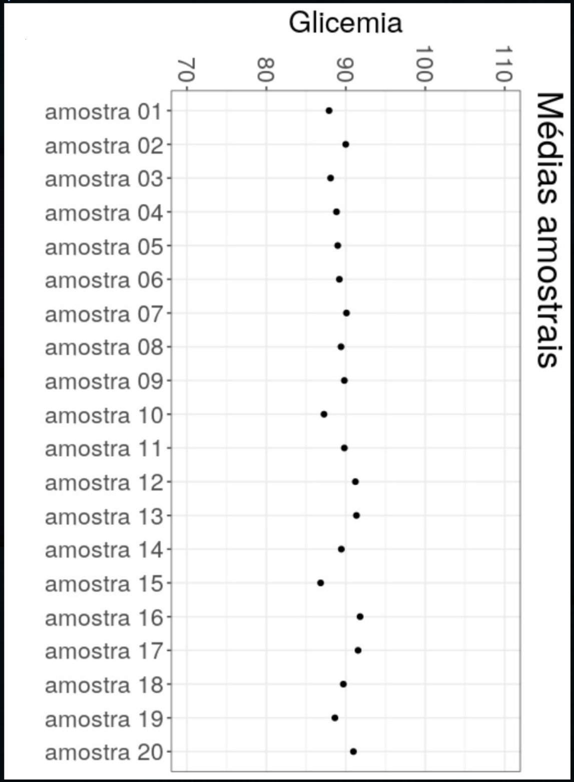
Com 20 amostras

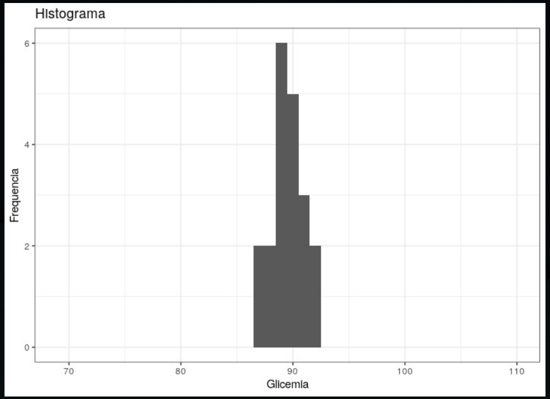




Com 20 amostras



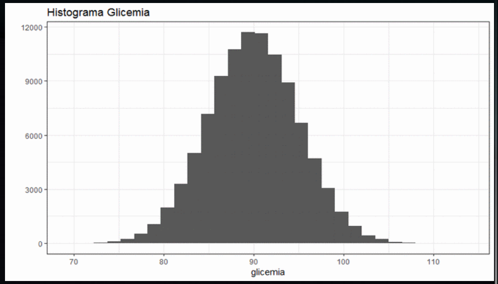




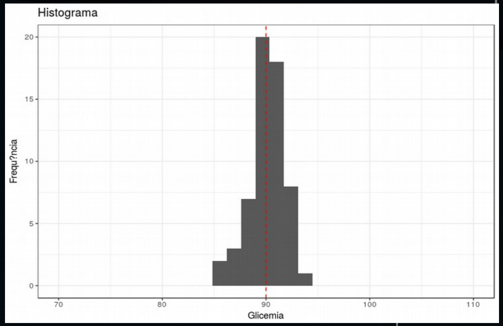
As médias amostrais têm distribuição normal

Do Teorema do Limite Central

População



Médias amostrais

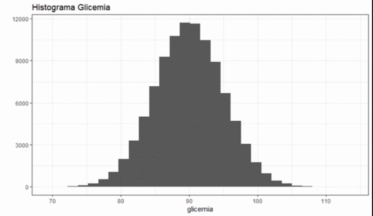


É possível calcular as médias amostrais prováveis que podem ser obtidas a partir de uma população com média populacional conhecida

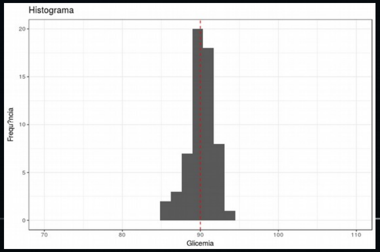
Do Teorema do Limite Central

De maneira análoga, é possível calcular as médias populacionais prováveis que podem ter gerado uma amostra com dada média amostral

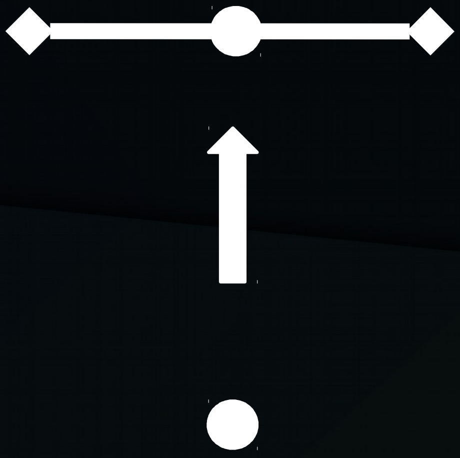
População



Médias amostrais



Intervalo de confiança



Médias Populacionais mais prováveis

Média amostral

Do Teorema do Limite Central

Intervalo de confiança:

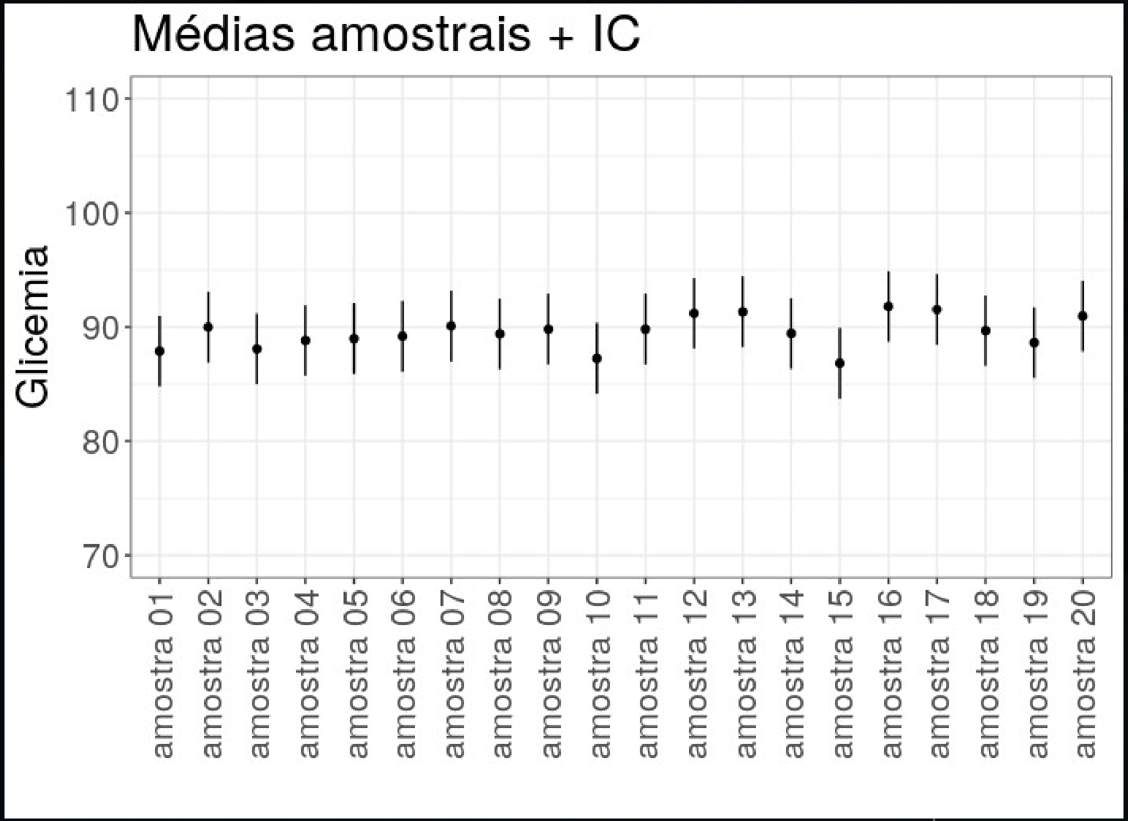
(Para 95% de confiança)

Limite inferior: $\bar{x} - 1{,}96 * \sigma_{\bar{x}}$

Limite superior: $\bar{x} + 1{,}96 * \sigma_{\bar{x}}$

Temos 95% de confiança de que a média populacional está entre estes dois valores

Com Intervalo de Confiança (95%)

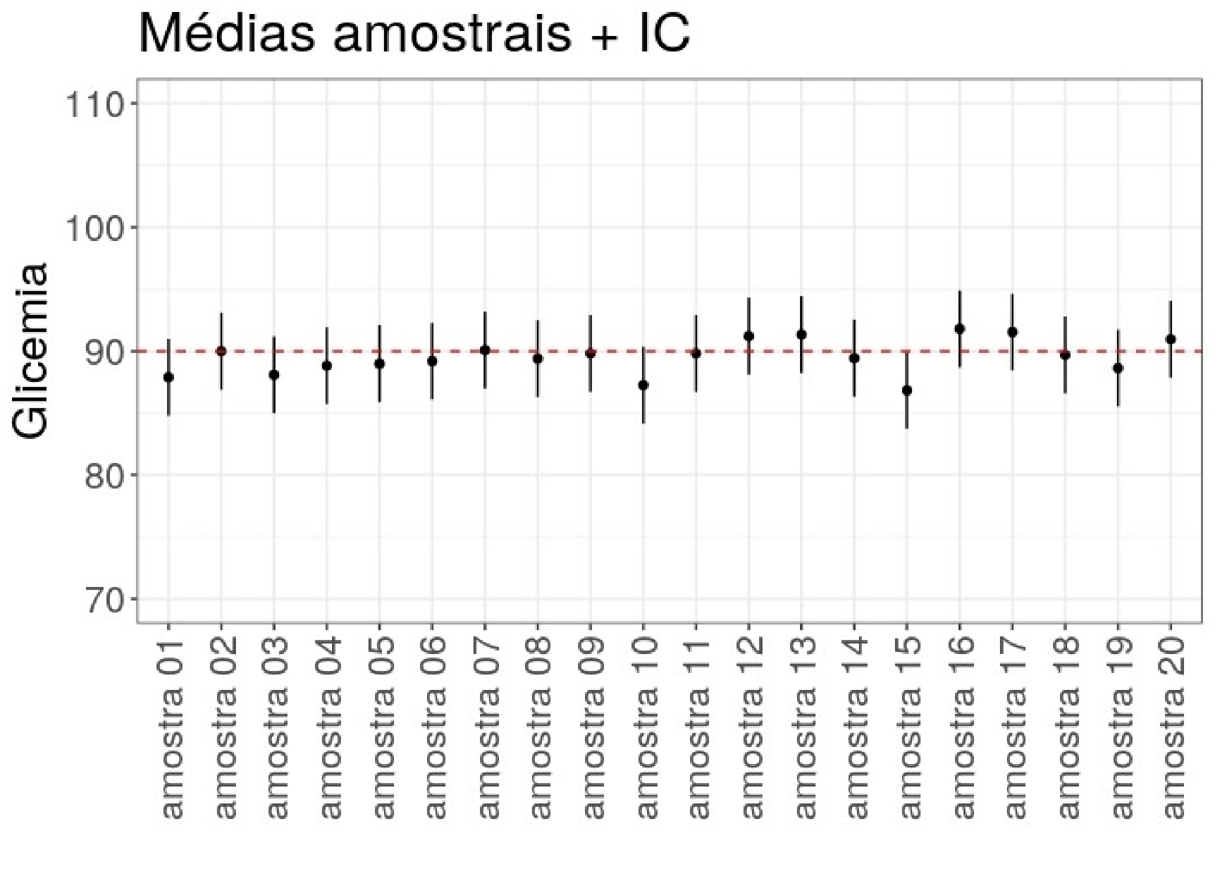


Para cada amostra, calculamos IC:

Lim sup = $\bar{x} + 1{,}96\ \sigma_{\bar{x}} =$
$\bar{x} + 1{,}96\ (\sigma / \sqrt{n})$

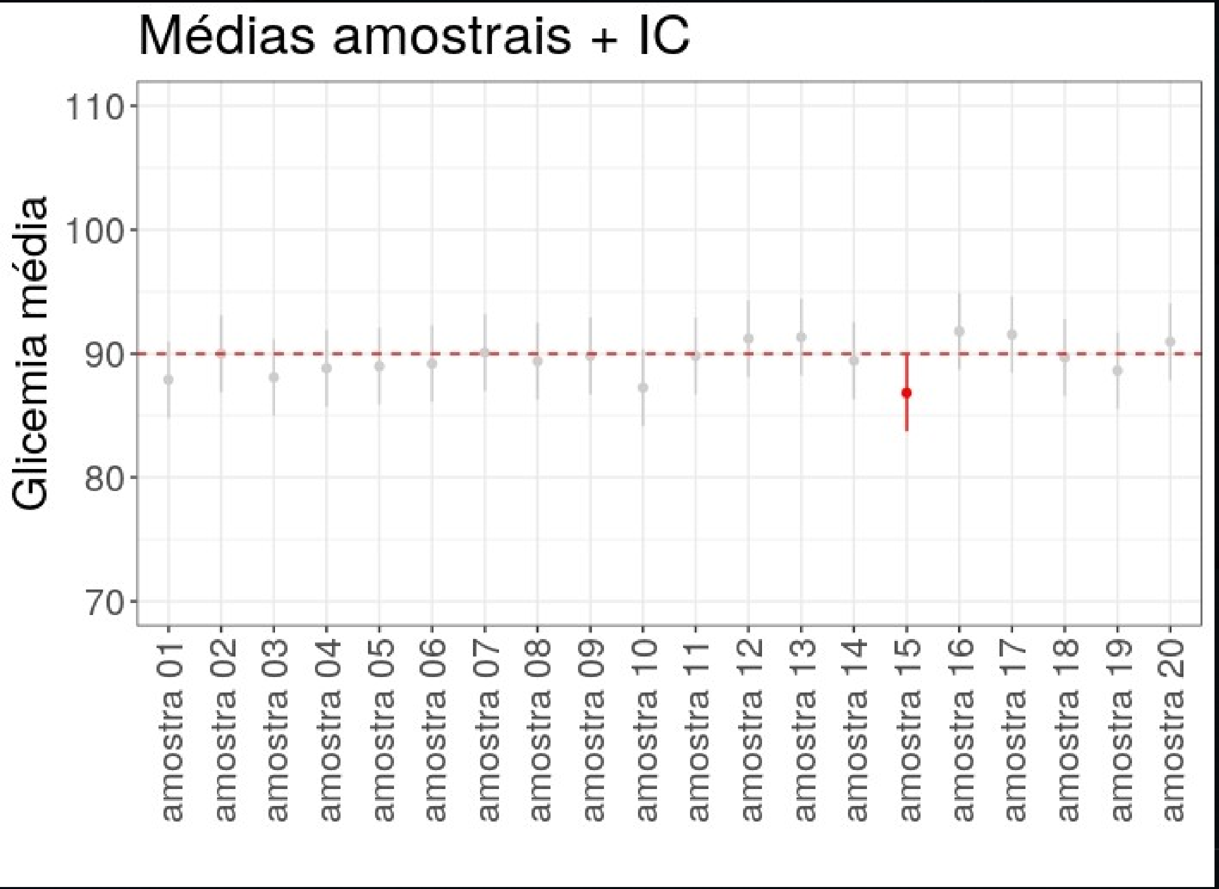
Lim inf = $\bar{x} – 1{,}96\ \sigma_{\bar{x}} =$
$\bar{x} - 1{,}96\ (\sigma / \sqrt{n})$

Qual era μ ?

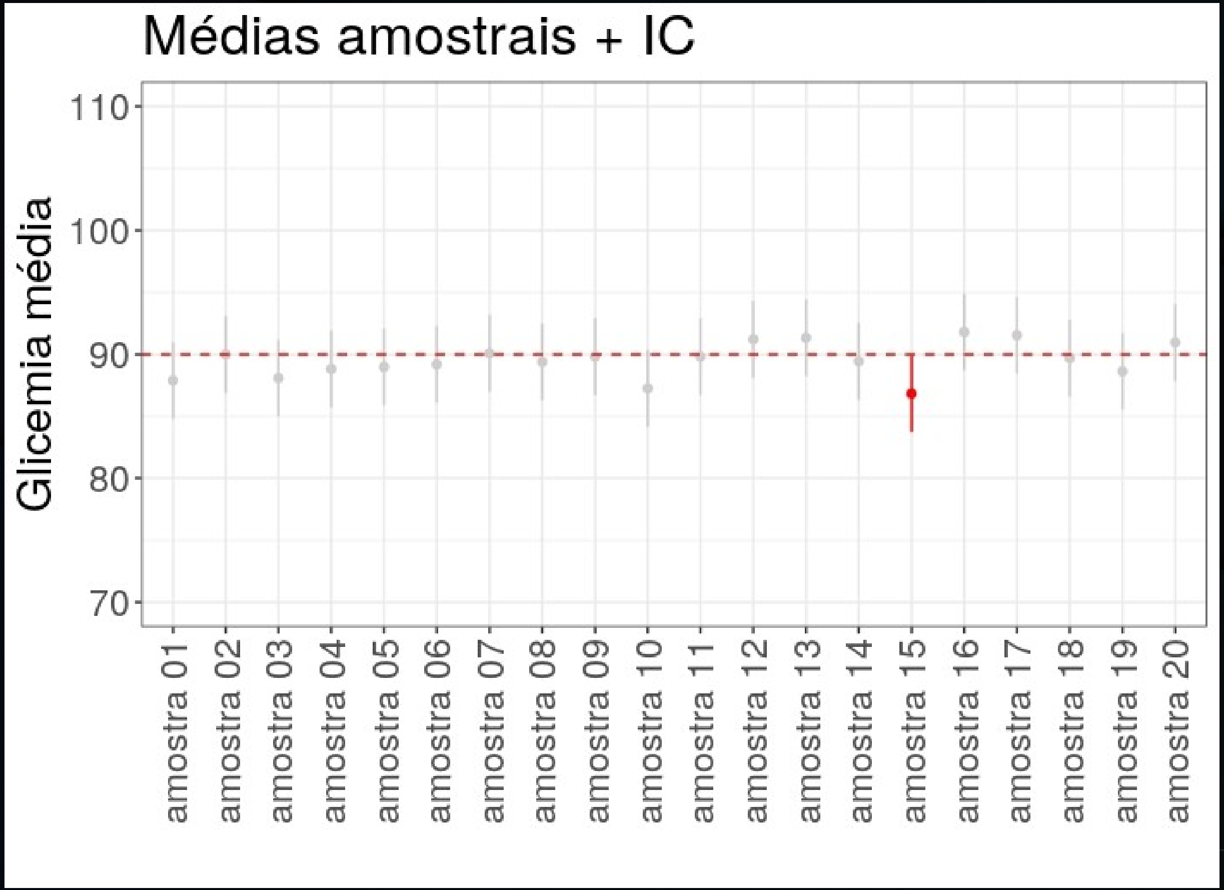


Médias amostrais + IC
Glicemia
110
100
90
80
70
amostra 01
amostra 02
amostra 03
amostra 04
amostra 05
amostra 06
amostra 07
amostra 08
amostra 09
amostra 10
amostra 11
amostra 12
amostra 13
amostra 14
amostra 15
amostra 16
amostra 17
amostra 18
amostra 19
amostra 20

Alguém errou?



Alguém errou?

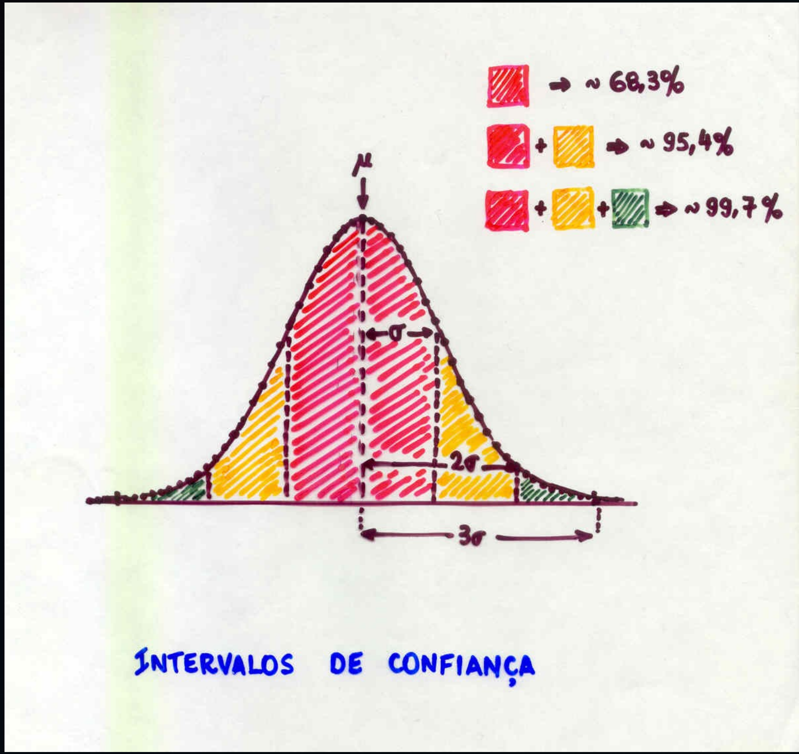


95% de confiança
Logo, P(erro) = 5% , do IC não incluir a média verdadeira.

5% = 1 / 20

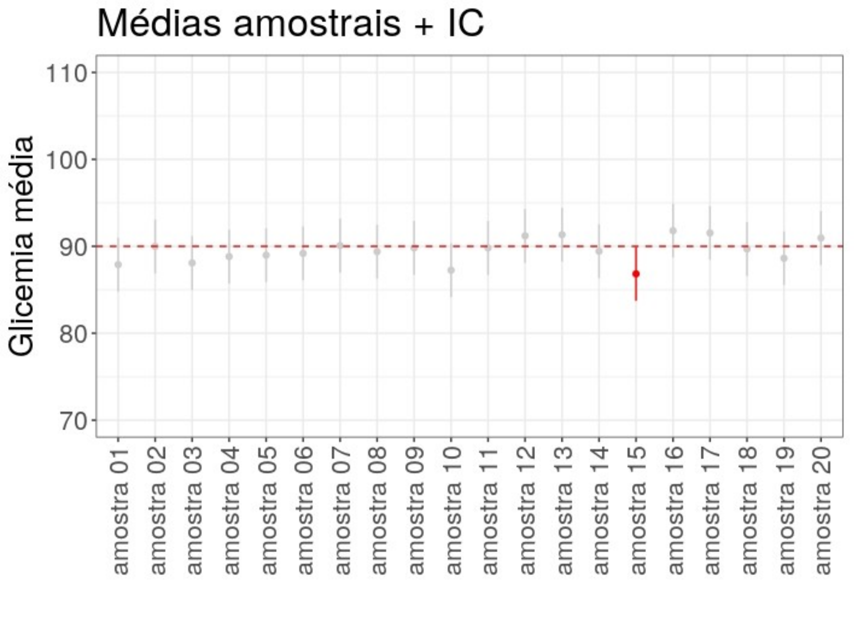
Ou seja, é esperado que 1 experimento a cada 20 erre.

Se quiser diminuir P(erro)

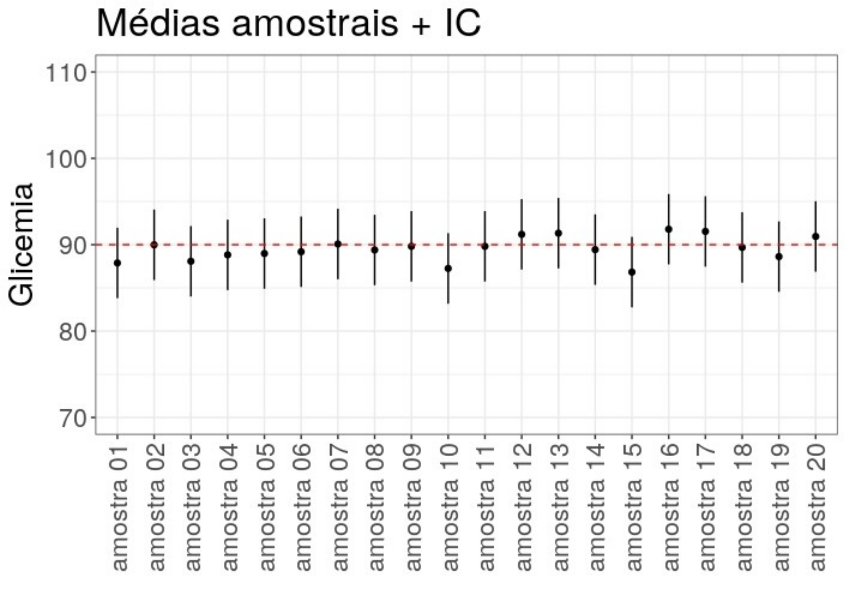


Confiança	P(erro)	Valor (z)
90%	10%	1.64
95%	**5%**	**1.96**
98%	2%	2.33
99%	1%	2.58

Com 99% confiança



IC (**95%**):
Lim sup = $\bar{x}$ + **1,96** (σ / √n)
Lim inf = $\bar{x}$ - **1,96** (σ / √n)

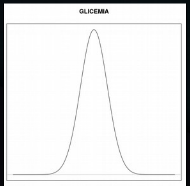


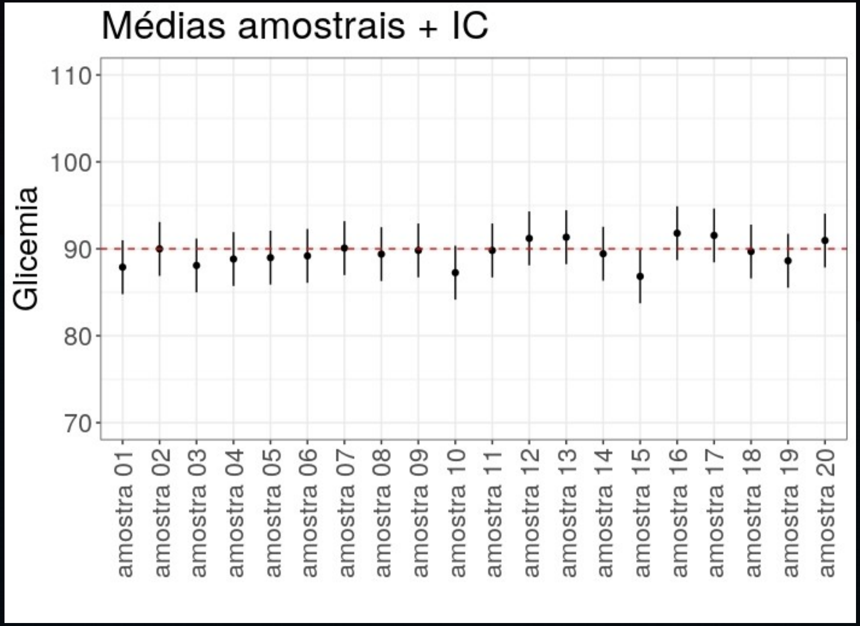
IC (**99%**):
Lim sup = $\bar{x}$ + **2,58** (σ / √n)
Lim inf = $\bar{x}$ – **2,58** (σ / √n)

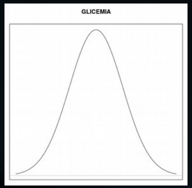
Impacto do σ

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

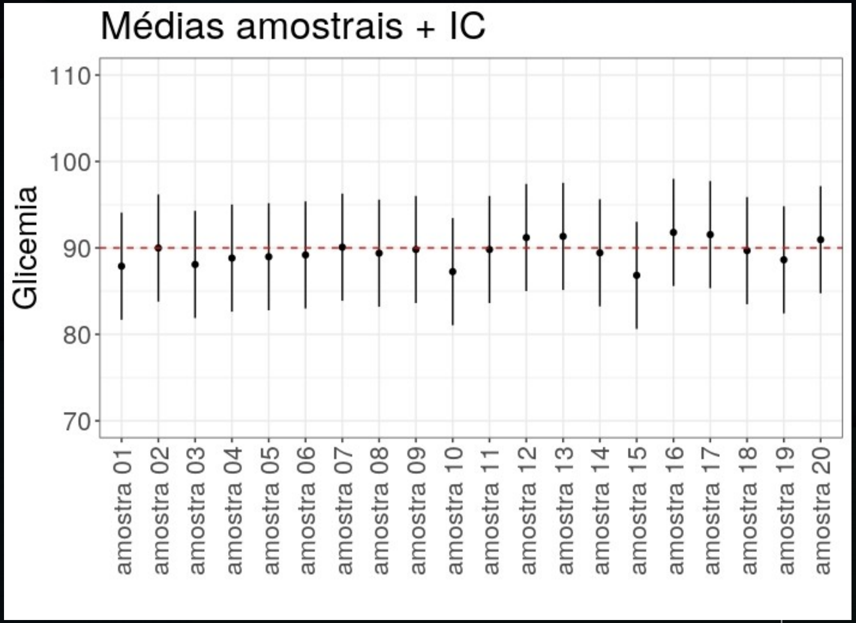
População 1: $\sigma = 5$







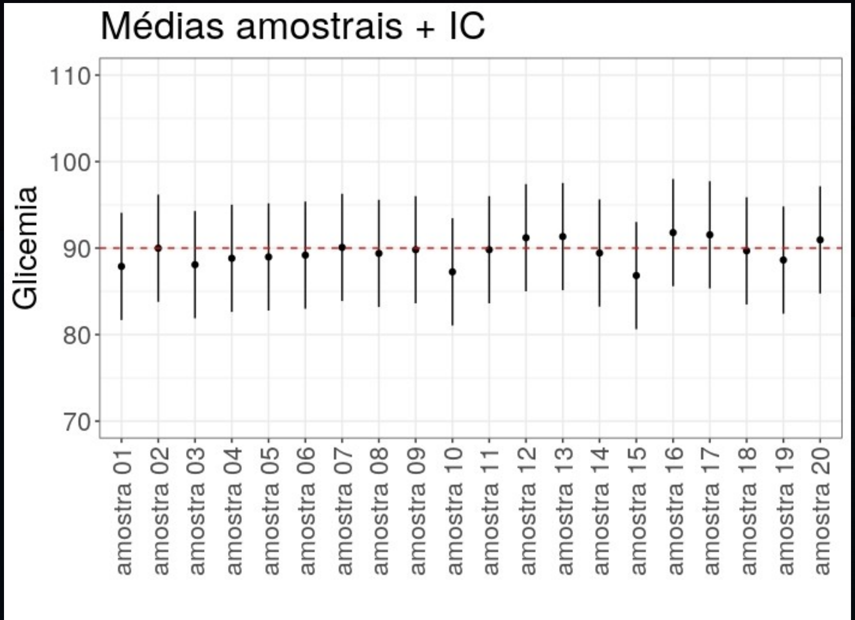
População 2: $\sigma = 10$



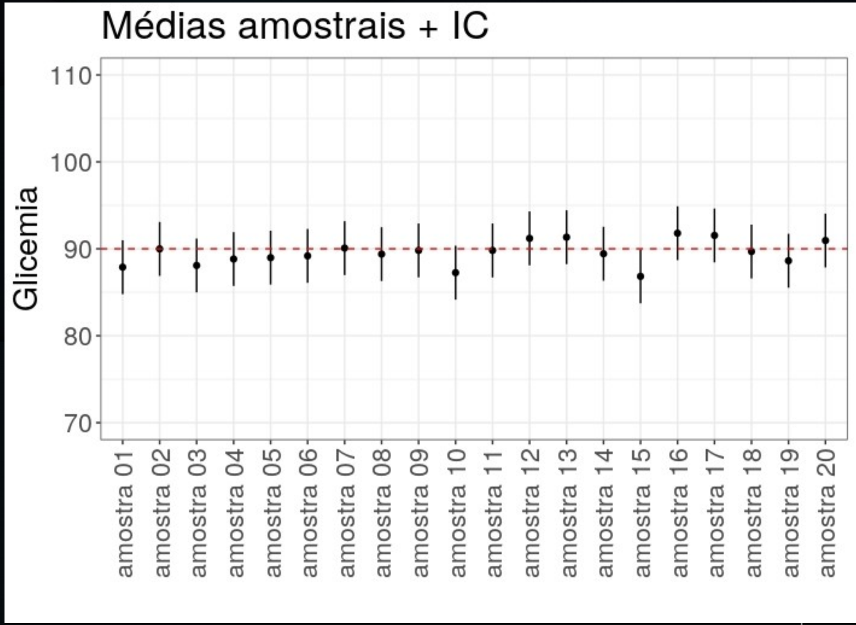
Impacto do tamanho da amostra

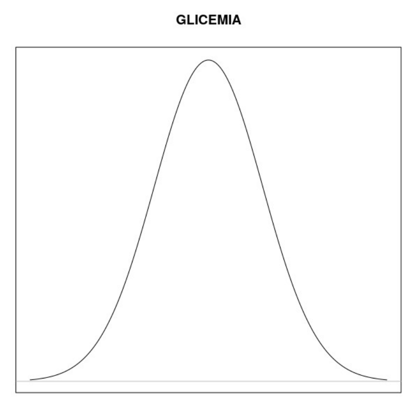
$$\varepsilon = 1{,}96\,\frac{\sigma}{\sqrt{n}}$$

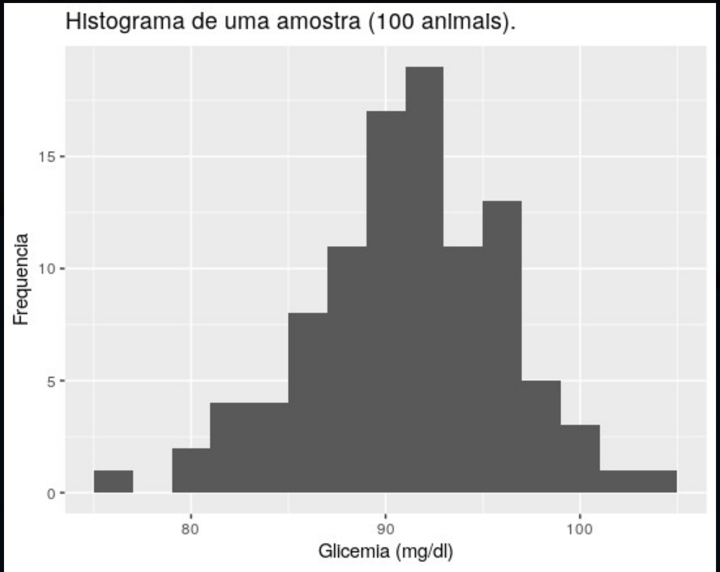
Amostras com n = 25



Amostras com n = 100







Exemplo
População: μ = ?

Amostra: $\bar{x}$ = 90,71

$IC(\mu,95\%) = \bar{x} \pm 1,96\ (\sigma / \sqrt{n})$

$IC(\mu,95\%)$ = 89,73 ; 91,69

Sei quanto é a média? Não.

Mas sei que ela deve estar nessa faixa de valores.

Legal, mas...

- Utilizamos $\bar{\mathbf{x}}$ para estimar **μ**
- Mas pra isso, precisamos de **σ**, que é o desvio-padrão populacional
- E o que fazemos se não conhecemos **σ** ? Situação muito comum

$$IC(\mu,95\%) = \bar{x} \pm 1{,}96\ (\sigma\ /\ \sqrt{n})$$

Sem saber **σ**

Duas soluções:

1 – Se o tamanho da amostra for grande, pode usar o *s* (desvio-padrão da amostra), pois nesses casos: $s \sim \sigma$

- O que é tamanho grande?

  Uns dizem 100 observações, outros 300. É mais ou menos daí pra cima.

2 – Usar a distribuição t de Student (ao invés da distribuição normal)

- Funciona pra qualquer tamanho, quando você não conhece a variância
- Na prática, todo mundo usa essa, que é mais prático.

Distribuição t de Student

William Gosset:
Mestre cervejeiro da Guinness
(também era estatístico e químico)

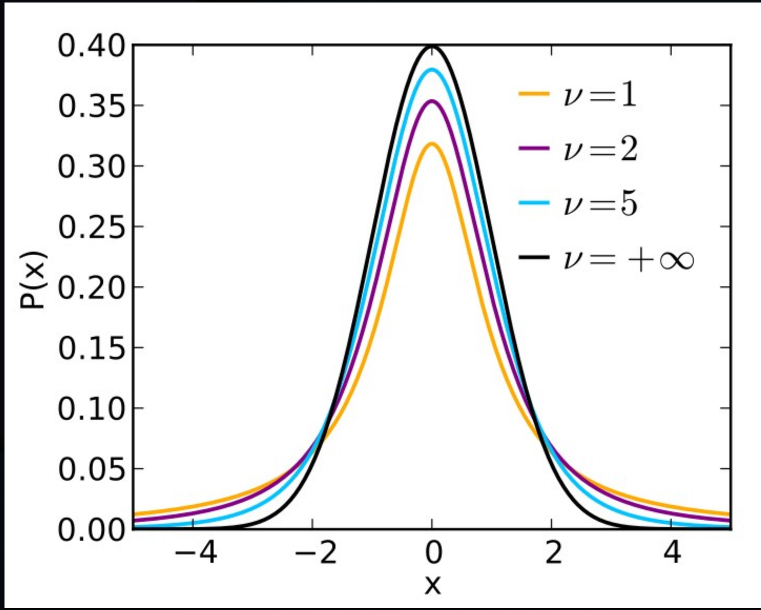
Pra fazer o controle de qualidade da matéria-prima da cerveja, trabalhava com amostras pequenas.

Descobriu que, mesmo quando a variância era desconhecida, as médias amostrais se distribuíam de uma maneira específica.

Publicou seus achados usando o pseudônimo de “Student”, regra da Guinness.

Essa distribuição ficou conhecida como t de Student.

Distribuição t de Student



$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

Fórmula macabra

Parece uma normal né?

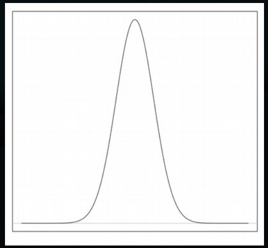
É como se a distribuição t de Student fosse quaaase uma normal, mas ajustando um pouco a curva, pra compensar a incerteza que temos sobre **σ**.

Quanto maior o tamanho da amostra, mais informação temos, e melhoramos nossa estimativa de **σ**.
Então quando n é muito grande, a distribuição t de Student fica quase igual uma normal.

Usando o mesmo raciocínio

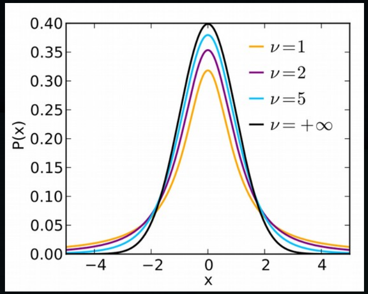
Usando a normal (TLC)

$$IC(\mu,95\%) = \bar{x} \pm 1{,}96 * (\sigma / \sqrt{n})$$



Usando a distribuição t de Student

$$IC(\mu;95\%) = \bar{x} \pm t_{95\%} * (s / \sqrt{n})$$



Depende do tamanho da amostra

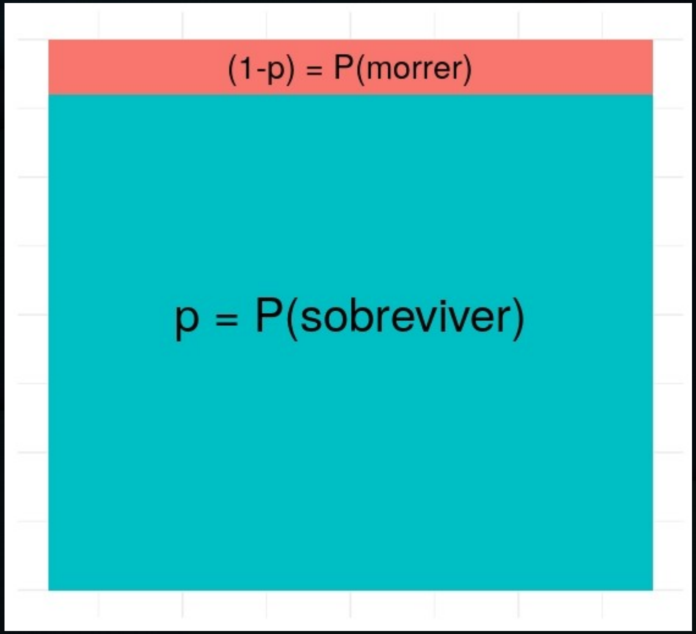
Distribuição t de Student

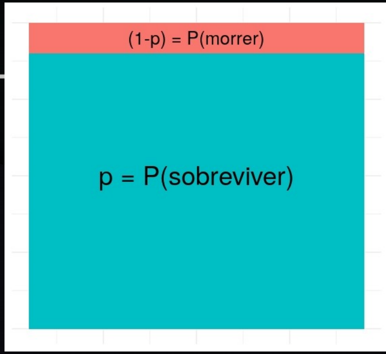
William Gosset:
Mestre cervejeiro da Guinness
(também era estatístico e químico)

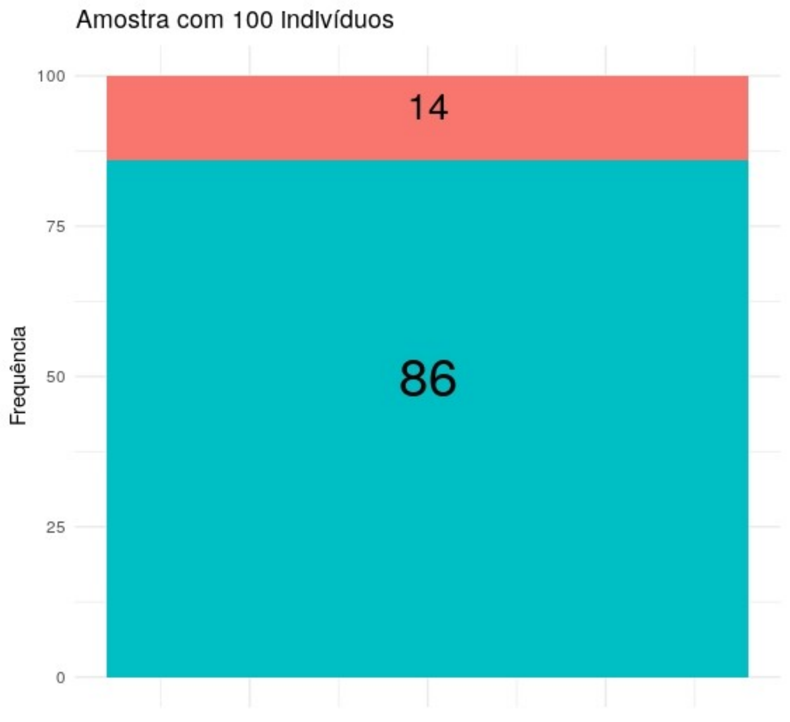
Variáveis qualitativas dicotômicas
Estimativa de proporção

Probabilidade p de um animal sobreviver à uma dada doença

Como estimar p, a partir de uma amostra?







Exemplo
População: p = ?

Amostra: $\hat{p}$ = 86%

Animal	Desfecho
1	Morreu
2	Sobreviveu
3	Morreu
...	...
100	Sobreviveu

Animal	Desfecho
1	Morreu
2	Sobreviveu
3	Morreu
...	...
100	Sobreviveu

Cada observação é um experimento de Bernoulli

Distribuição de Bernoulli:

$\mu = p$
$\sigma^2 = p(1-p)$
$\sigma = \sqrt{p(1-p)}$

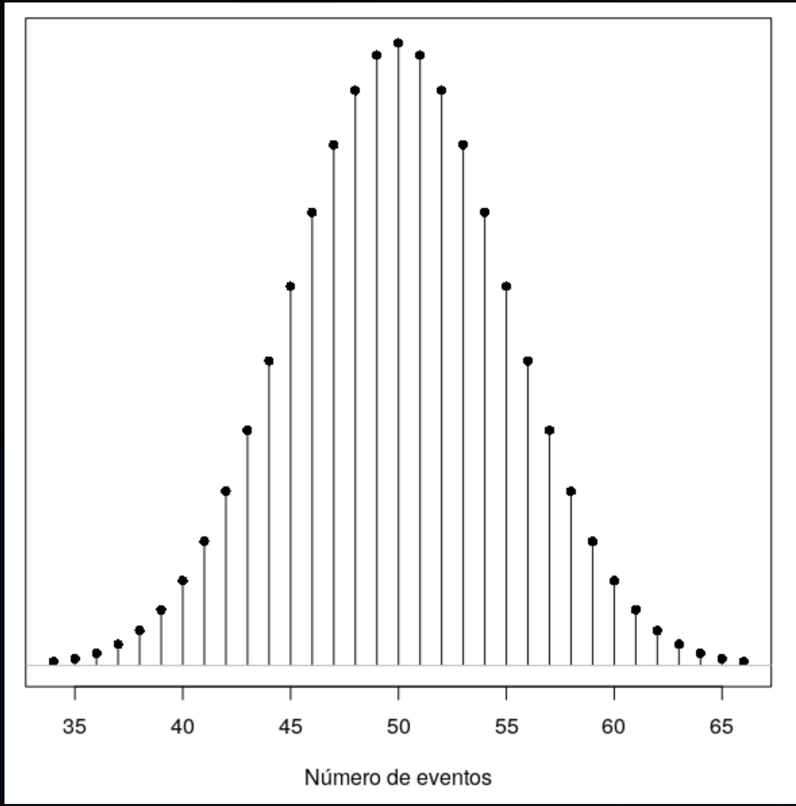
Juntar n experimentos Bernoulli → Binomial

Desfecho	Total
Morreu	14
Sobreviveu	86

Distribuição Binomial:

$\mu = np$
$\sigma^2 = np(1-p)$
$\sigma = \sqrt{np(1-p)}$

Binomial → Normal *



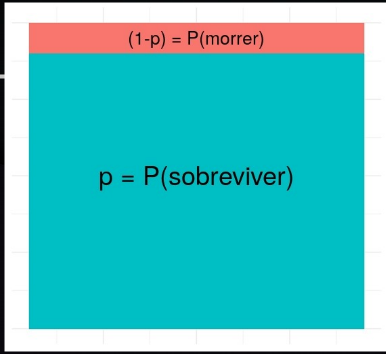
Usando a normal (TLC)

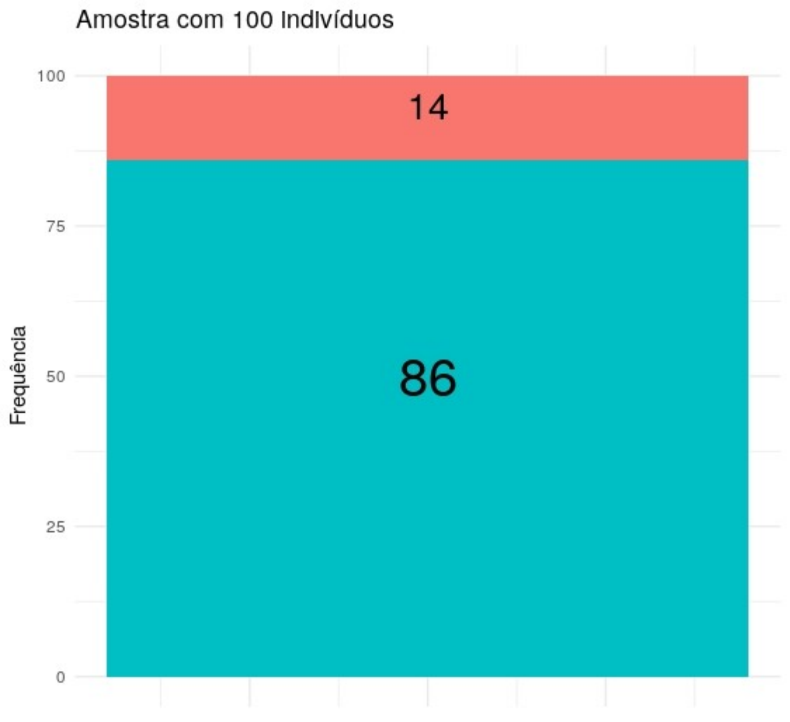
$$IC(\mu,95\%) = \bar{x} \pm 1{,}96 * \frac{\sigma}{\sqrt{n}}$$

Substituindo média e desvio-padrão

$$IC(p;95\%) = \hat{p} \pm 1{,}96 * \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

* Quando np > 5 e n(1 – p) > 5 (Morettin)





Exemplo

População: p = ?

Amostra: $\hat{p}$ = 86%

$$IC(p;95\%) = \hat{p} \pm 1,96 * \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

$IC(p;95\%)$ = 78% ; 92%

Sei a probabilidade exata de sobrevivência?
Não.

Mas sei que ela deve estar nessa faixa de valores.

Obrigado