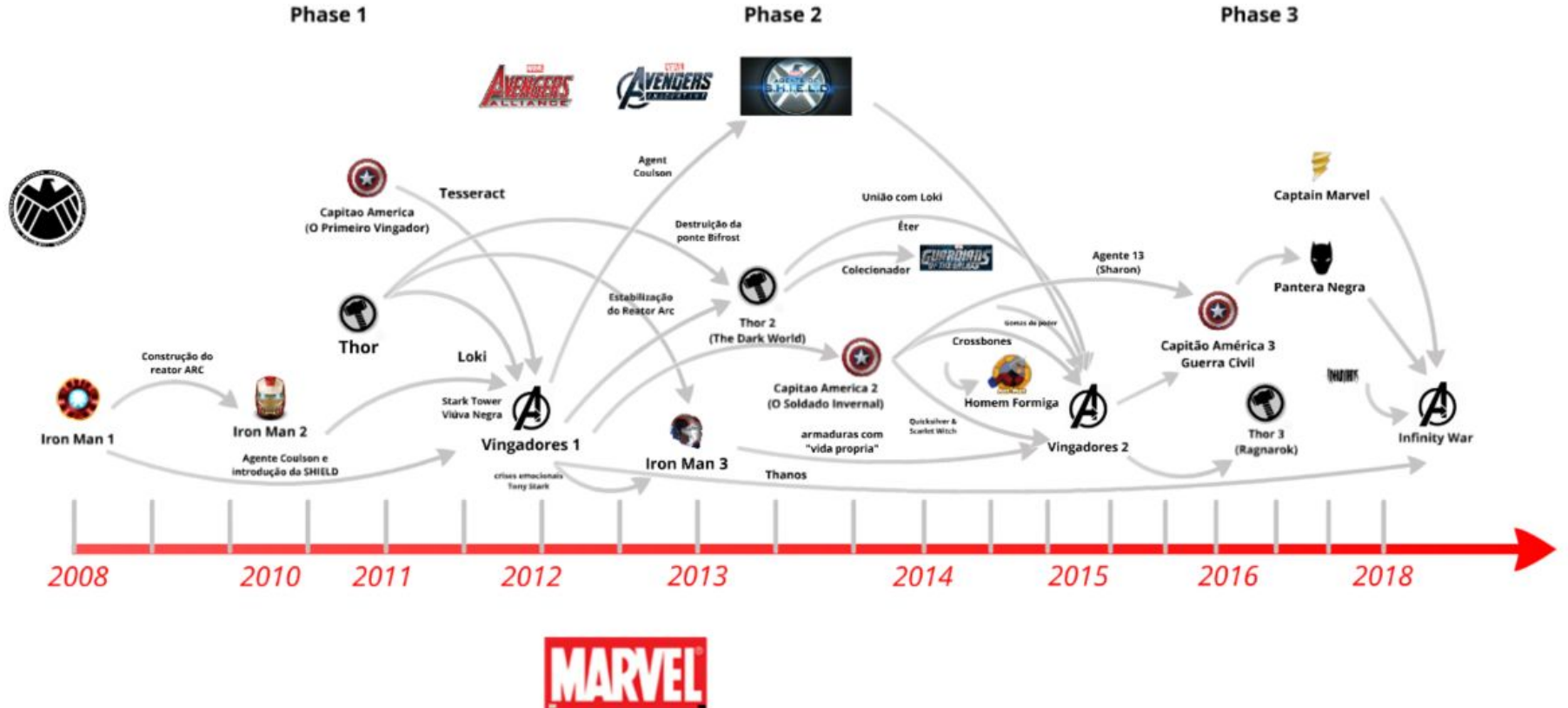
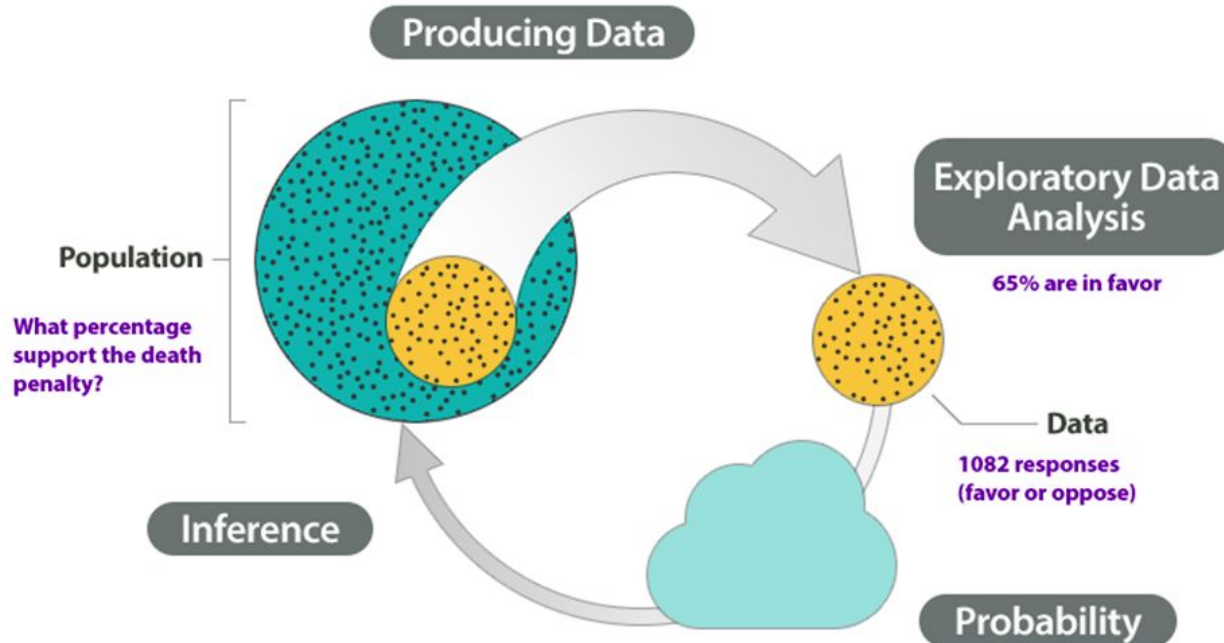


INFERÊNCIA

Project Avengers



INFERÊNCIA ESTATÍSTICA

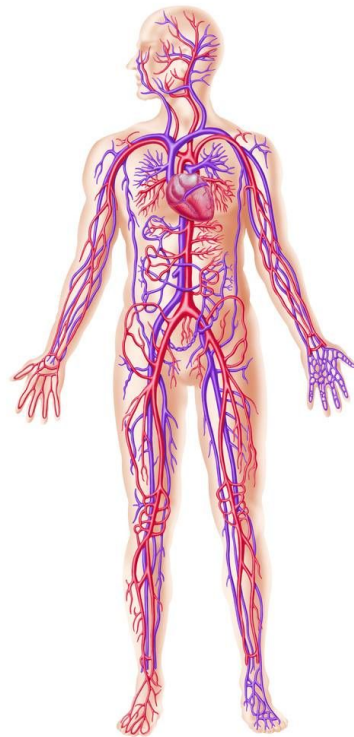


Conclusion: we can be 95% sure that the population percentage is within 3% of 65% (i.e. between 62% and 68%).

INFERÊNCIA ESTATÍSTICA

Corpo humano: 5-6 litros de sangue

Teste de sangue: 3-10mL (0,05% - 0,2%)



INFERÊNCIA ESTATÍSTICA

Brasil: 146,4MM votantes

Pesquisa Eleitorais: ~2.000 (0,001366%)



INFERÊNCIA ESTATÍSTICA

**Prós e contras de pesquisas de opinião pública - Jason
Robert Jaffe**

<https://www.youtube.com/watch?v=ubR8rEgSZSU>

AMOSTRAGEM

Não probabilística

- Conveniência
- Voluntaria
- Snowball
- ...

Probabilística

- Aleatória simples
- Estratificada
- Cluster
- ...

CONVENIÊNCIA



VOLUNTÁRIA

Google Form to Sheets

QUESTIONS RESPONSES

Google Form to Sheets - Test

Used for the Tuts+ tutorial.

What do you want to learn in this course? *

☐ Web design

☐ Public speaking

☐ Graphic design

Question

☐ Option 1

☐ Add option or ADD "OTHER"

Short answer

Paragraph

Multiple choice

Checkboxes

Dropdown

Linear scale

Multiple choice grid

Date

Time



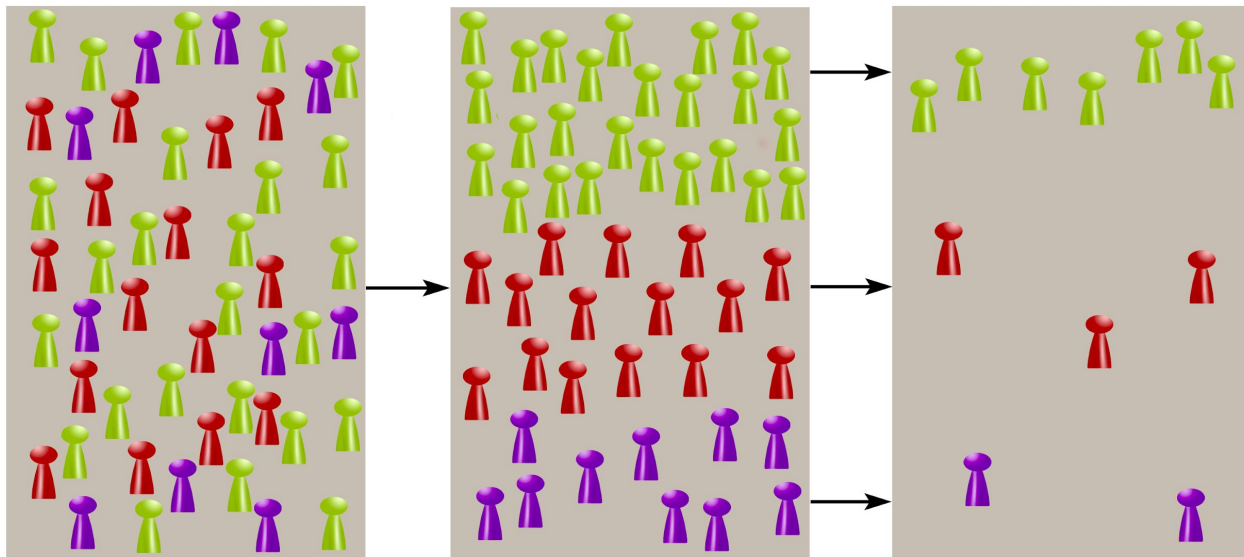
BOLA DE NEVE (SNOWBALL)



ALEATÓRIA SIMPLES



ESTRATIFICADA



CLUSTER



Considere as seguintes afirmações:

I- População: alunos de um curso. A professora escreve o nome de todos os alunos em pedaços de papel e os coloca em uma caixa. Depois de misturá-los, o professor sorteia 10 nomes.

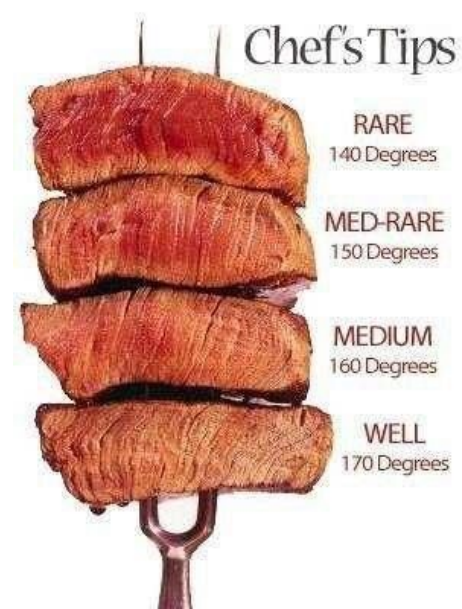
II- População: alunos de um curso. Um professor selecionará uma amostra de alunos para estimar a proporção de alunos que fizeram o dever de casa. 1) Alunos com notas baixas têm duas vezes mais chances de serem selecionados, comparados a um aluno com notas altas; 2) os alunos sorteados aleatoriamente na turma anterior não participam deste sorteio.

III- População: pessoas em fila para a próxima sessão de cinema. Um administrador de uma sala de cinema realiza uma pesquisa para estimar a proporção de pessoas nesta fila que desejam comprar pipoca. Ele seleciona cada 10 pessoas na fila. A primeira pessoa selecionada é escolhida aleatoriamente entre as 10 primeiras da fila.

Os métodos de amostragem descritos acima são, respectivamente:

- a) Amostra aleatória simples; Amostra não probabilística; Amostra probabilística.
- b) Amostra aleatória simples; Amostra probabilística; Amostra não probabilística.
- c) Amostra aleatória simples; Amostra não probabilística; Amostra não probabilística.
- d) Amostra não probabilística; Amostra não probabilística; Amostra probabilística.
- e) Amostra não probabilística; Amostra probabilística; Amostra não probabilística.

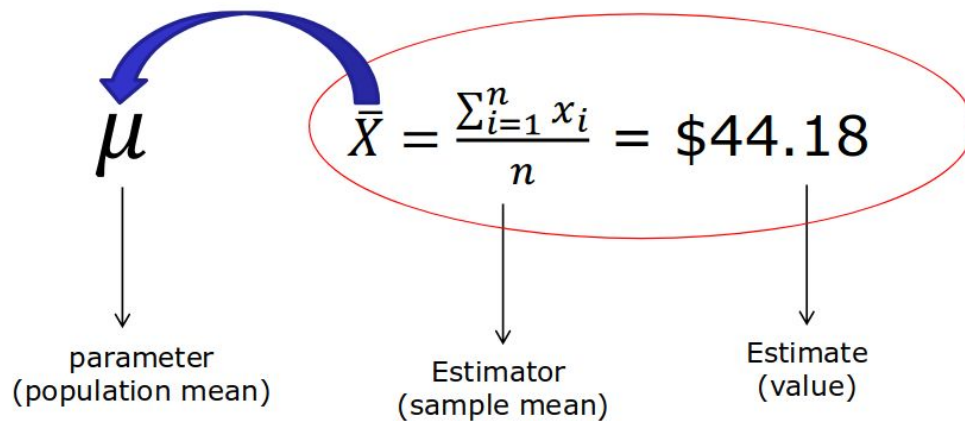
ESTIMADORES



ESTIMADORES

Parâmetro	População	Amostra
Média	μ	\bar{x}
Variância	σ^2	s^2
Proporção	p	\bar{p}
Diferença	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$

ESTIMADORES



TANQUES ALEMÃES

Em 1943, você trabalhava para as Forças Aliadas.
Sete tanques alemães acabaram de ser capturados. Os tanques do exército alemão têm um número de série começando com 1 e terminando com n, onde n=número de tanques do exército alemão.

Os números de série dos tanques capturados eram:

41, 105, 191, 271, 290, 315 e 372.

Sua missão:

Estime o número de tanques do Exército Alemão com base nos números de série desta amostra.



TANQUES ALEMÃES

Month	Statistical estimate	Intelligence estimate	German records
June 1940	169	1,000	122
June 1941	244	1,550	271
August 1942	327	1,550	342

https://en.wikipedia.org/wiki/German_tank_problem

Technology

Technology blog

Charles Arthur

@charlesarthur

Wed 8 Oct 2008 11:41 BST



4 8

Why iPhones are just like German tanks

You might not see the similarity at once, but the latest estimates for iPhone sales come straight from World War II

(Photo from Flickr by [PhotosNormandie](#). Some rights reserved.)

You'll have noticed the PDA blog's report about the [latest estimates of how many iPhones have been sold](#) - a solid 9.1m and counting.

But you might have wondered how the process of simply asking people for their phone's serial number, IMEI number and date of purchase would allow someone to make such a specific calculation of the number sold (which, according to the [spreadsheet](#), was 9,190,680).

The answer: tanks. Well, the same process that allowed the Allies to calculate how many tanks the Nazis were producing during World War II. You can [read Gavyn Davies's full-length account](#), or we'll give you the extracted version here.

It goes like this:

By 1941-42, the allies knew that US and even British tanks had been technically superior to German Panzer tanks in combat, but they were worried about the capabilities of the new marks IV and V. More troubling, they had really very little idea of how many tanks the enemy was capable of producing in a year. Without this information, they were unsure whether any invasion of the continent on the western front could succeed.

They could try spying on factories, or count tanks in the field. This was easier. Even better: get the serial numbers from the tanks, and estimate from this how many were being built.

most viewed



Live Toronto suspect named as Alek Minassian after van hits pedestrians - as it happened



Toronto van incident in which 10 pedestrians died appeared deliberate, say police



Finland to end basic income trial after two years

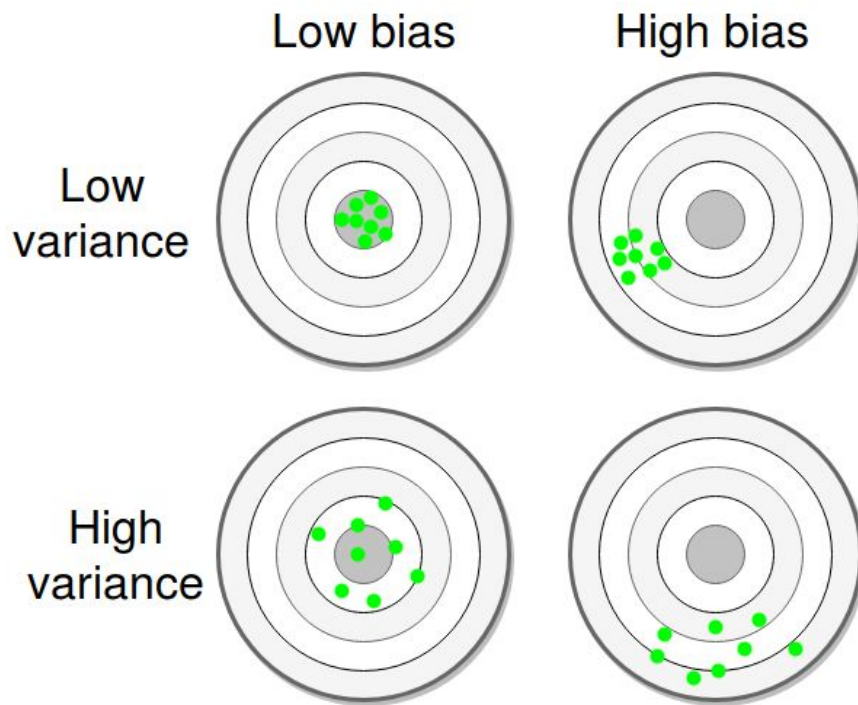


Why are the Bushes, Clintons, Obamas and Melania smiling so broadly at a funeral?



Verne Troyer's tragic death underlines the harm Mini-Me caused people with dwarfism | Eugene Grant
[Eugene Grant](#)

VIÉS E VARIÂNCIA



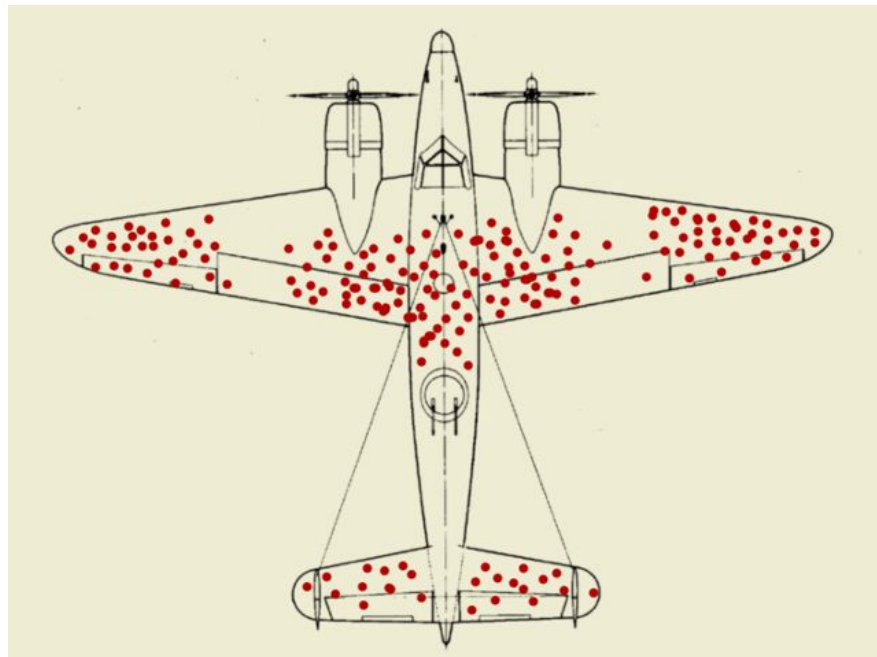
VIÉS

De cada **100 aviadores**, 45 foram mortos, 6 ficaram gravemente feridos, 8 tornaram-se prisioneiros de guerra e apenas **41 escaparam ilesos** – pelo menos fisicamente. Dos que voavam no início da guerra, **apenas 10% sobreviveram.**”



<https://www.dgsiegel.net/talks/the-bullet-hole-misconception>

VIÉS



VIÉS

You are missing something! - Survivorship bias

<https://www.youtube.com/watch?v=ZyLVlvBidIA>

VIÉS

Can tech be biased? CNN Money

https://www.youtube.com/watch?v=gTtbseMYm_s

You are missing something! - Survivorship bias

<https://www.youtube.com/watch?v=ZyLVlvBidIA>

VARIÂNCIA

VARIÂNCIA

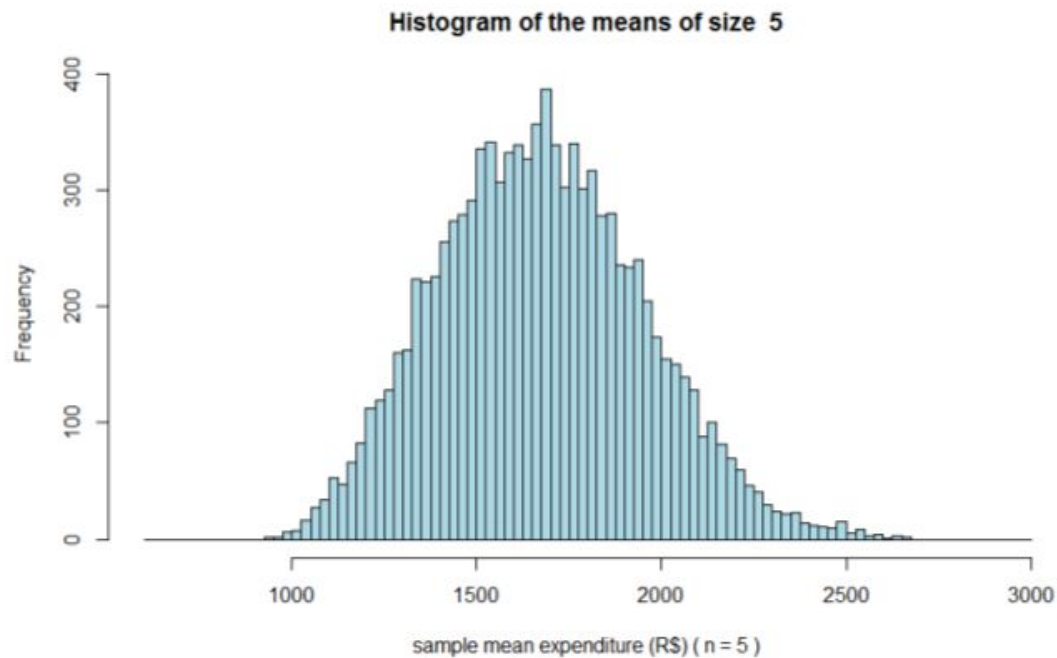
Código em R

```
M = 100.000           # number of samples
Xbar = rep(0,M)        # vector that stores the means
n = 5                 #sample size
X = 1000:3000

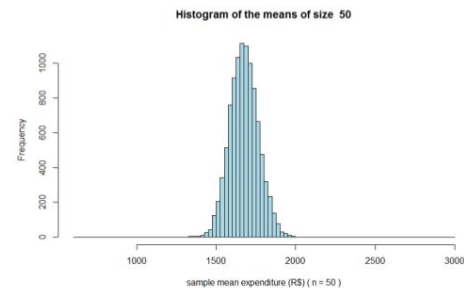
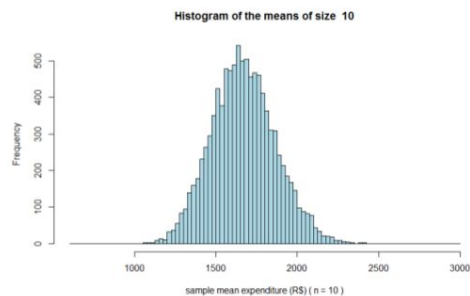
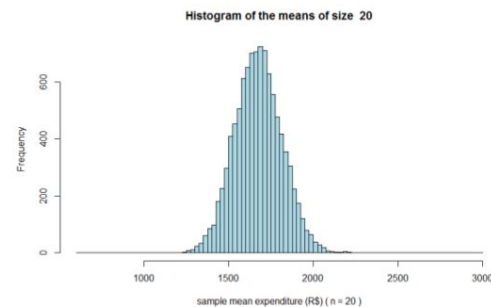
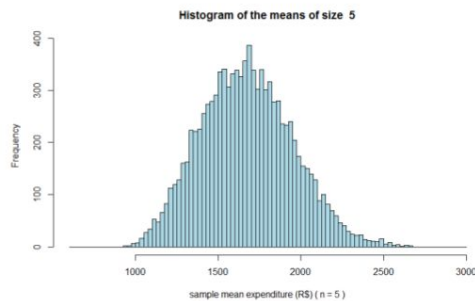
for (i in 1:M)
{
  one_sample = sample(X,size=n,replace=TRUE)
  Xbar[i] = mean(one_sample)
}

hist(Xbar, xlim = c(1000,3000))
```

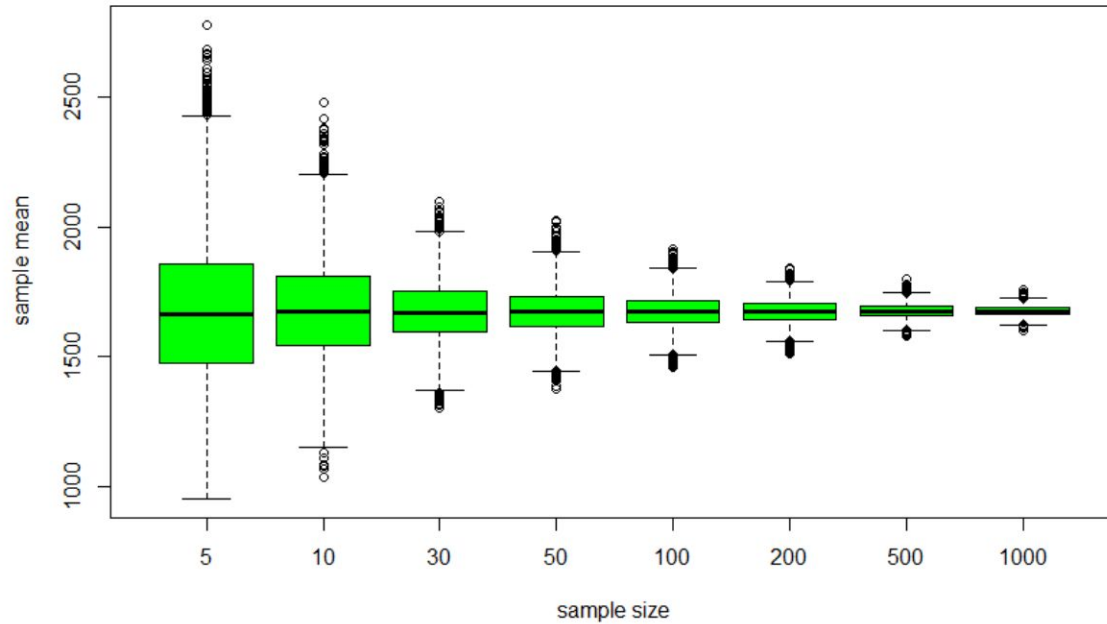
VARIÂNCIA



VARIÂNCIA



VARIÂNCIA



Source: prof Andre Samartini

TEOREMA DO LIMITE CENTRAL

TÁBUA DE GALTON

Galton Board

<https://www.youtube.com/watch?v=6YDHBFBVlvls>



TEOREMA DO LIMITE CENTRAL

“N” moedas

$$\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

	0	1	2	3	4
N = 1					
N = 2					
N = 3					
N = 4					

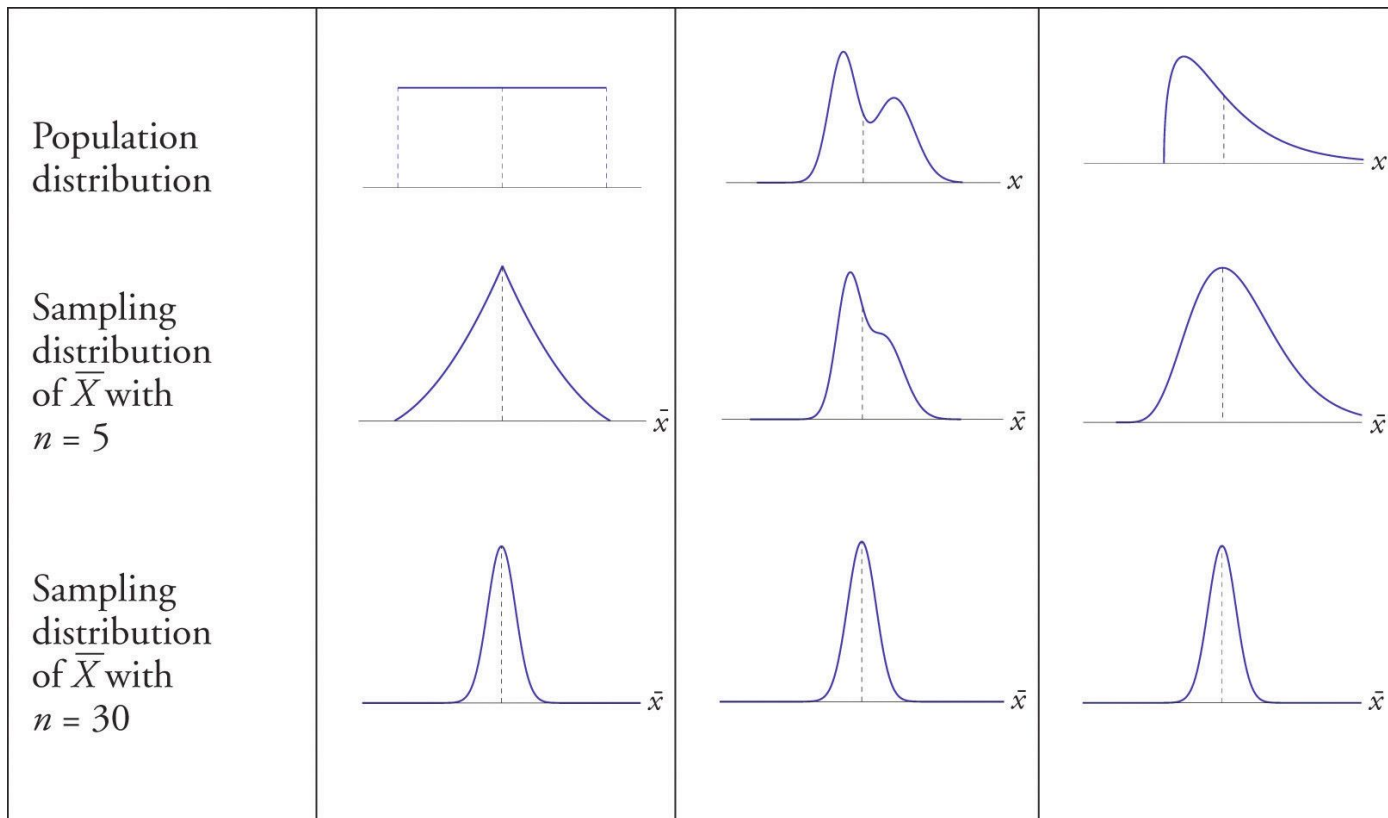
TEOREMA DO LIMITE CENTRAL

“N” moedas

$$\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

	0	1	2	3	4
N = 1	1	1			
N = 2	1	2	1		
N = 3	1	3	3	1	
N = 4	1	4	6	4	1

TEOREMA DO LIMITE CENTRAL



TEOREMA DO LIMITE CENTRAL

“Given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately **equal to the mean of the population**. Furthermore, all of the samples will follow an approximate normal distribution pattern...”

https://www.investopedia.com/terms/c/central_limit_theorem.asp

TEOREMA DO LIMITE CENTRAL

$$[x_1 + \dots + x_n]/n = \sum x_i/n \sim N(\mu, ?)$$

TEOREMA DO LIMITE CENTRAL

$$\begin{aligned}\text{Var}([x_1 + \dots + x_n]/n) &= 1/n^2 * \text{Var}([x_1 + \dots + x_n]) \\ &= 1/n^2 * [\text{Var}(x_1) + \dots + \text{Var}(x_n)] \\ &= 1/n^2 * [\sigma^2 + \dots + \sigma^2] \\ &= 1/n^2 * n[\sigma^2] \\ &= \sigma^2/n\end{aligned}$$

TEOREMA DO LIMITE CENTRAL

$$[x_1 + \dots + x_n]/n = \sum x_i/n \sim N(\mu, \sigma^2/n)$$



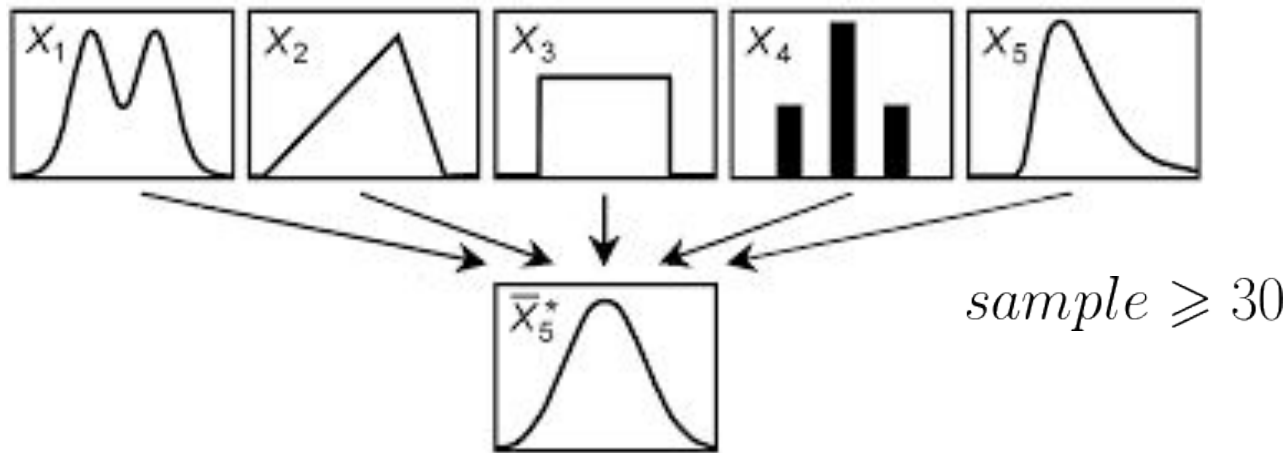
(Erro padrão)²

TEOREMA DO LIMITE CENTRAL

Bunnies, Dragons and the 'Normal' World: Central Limit Theorem | The New York Times

<https://www.youtube.com/watch?v=jvoxEYmQHNM>

TEOREMA DO LIMITE CENTRAL



$$\bar{x} \sim N \left(mean = \mu, sd = \frac{\sigma}{\sqrt{n}} \right)$$

TEOREMA DO LIMITE CENTRAL

Para estimar os gastos médios mensais com alimentação para os 10.000 touros da população, foi selecionada uma amostra aleatória de 250 touros e a média amostral foi de US\$ 27. Considere que o desvio padrão da população é de US\$ 3,00. Com base nestes resultados, qual é o **erro padrão** desta estimativa?

- A. 0.012
- B. 0.108
- C. 0.189
- D. 0.300
- E. 1.718

TEOREMA DO LIMITE CENTRAL

A variável X tem uma distribuição de Poisson. Qual das afirmações a seguir é verdadeira?

- a) Se retirarmos um número muito grande (>30) de amostras (de qualquer tamanho) de uma população, o formato do histograma dos valores das respectivas médias amostrais tenderá a ser distribuído normalmente.
- b) Se retirarmos um número muito grande de amostras (de tamanho >30) de uma população, o formato do histograma dos valores das respectivas médias amostrais tenderá a ser distribuído normalmente.
- c) Se retirarmos um grande número (> 30) de amostras (de qualquer tamanho) de uma população, o formato do histograma dos valores das respectivas médias amostrais tenderá a ter o formato do histograma da distribuição da população.
- d) Se retirarmos um grande número de amostras (de tamanho >30) de uma população, o formato do histograma dos valores das respectivas médias amostrais tenderá a ter o formato do histograma da distribuição da população.

TEOREMA DO LIMITE CENTRAL

A variável X tem uma distribuição de Poisson. Qual das afirmações a seguir é verdadeira?

a) Se retirarmos um número muito grande (>30) de amostras (de qualquer tamanho) de uma população, o formato do histograma dos valores das respectivas médias amostrais tenderá a ser distribuído normalmente.

b) Se retirarmos um número muito grande de amostras (de tamanho >30) de uma população, o formato do histograma dos valores das respectivas médias amostrais tenderá a ser distribuído normalmente.

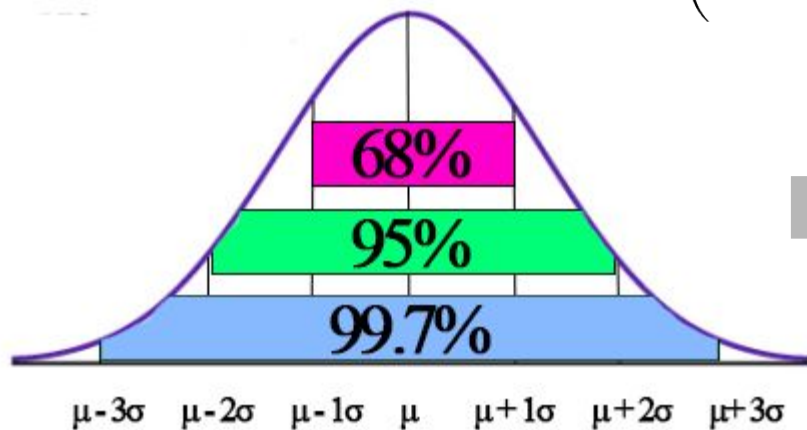
c) Se retirarmos um grande número (> 30) de amostras (de qualquer tamanho) de uma população, o formato do histograma dos valores das respectivas médias amostrais tenderá a ter o formato do histograma da distribuição da população.

d) Se retirarmos um grande número de amostras (de tamanho >30) de uma população, o formato do histograma dos valores das respectivas médias amostrais tenderá a ter o formato do histograma da distribuição da população.

INTERVALO DE CONFIANÇA

INTERVALO DE CONFIANÇA

$$\bar{x} \sim N \left(mean = \mu, sd = \frac{\sigma}{\sqrt{n}} \right)$$



Quantos “desvios padrão”?

$$\mu = \bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

A blue arrow points from the text 'Quantos “desvios padrão”?' to the $z_{\frac{\alpha}{2}}$ term in the formula, which is circled in blue. A red arrow points from the same text to the $\frac{\sigma}{\sqrt{n}}$ term, which is circled in red.

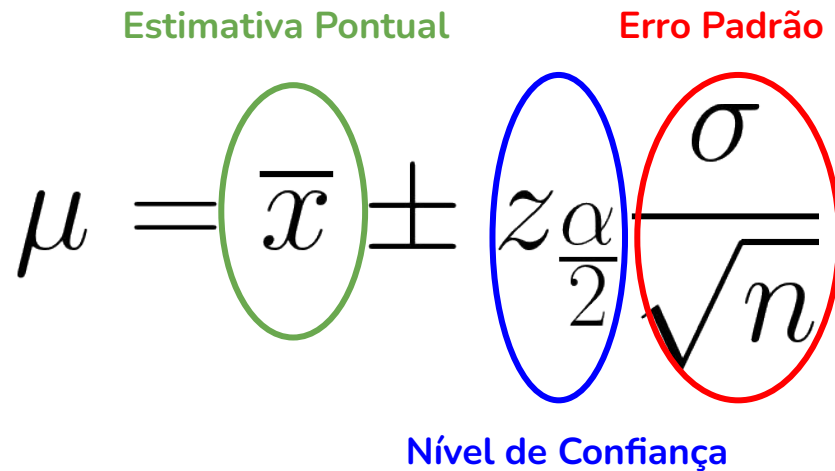
INTERVALO DE CONFIANÇA

Estimativa Pontual

Erro Padrão

$$\mu = \bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Nível de Confiança

The diagram shows the confidence interval formula with three color-coded annotations. A green oval encircles the sample mean \bar{x} , with the label 'Estimativa Pontual' above it. A blue oval encircles the critical value $z_{\frac{\alpha}{2}}$, with the label 'Nível de Confiança' below it. A red oval encircles the standard error term $\frac{\sigma}{\sqrt{n}}$, with the label 'Erro Padrão' above it.

INTERVALO DE CONFIANÇA

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Margem de Erro

Níveis de confiança	Z $\alpha/2$
90%	1,645
95%	1,960
99%	2,576

INTERVALO DE CONFIANÇA

Qual o valor de $Z_{\alpha/2}$ para os seguintes níveis de confiança:

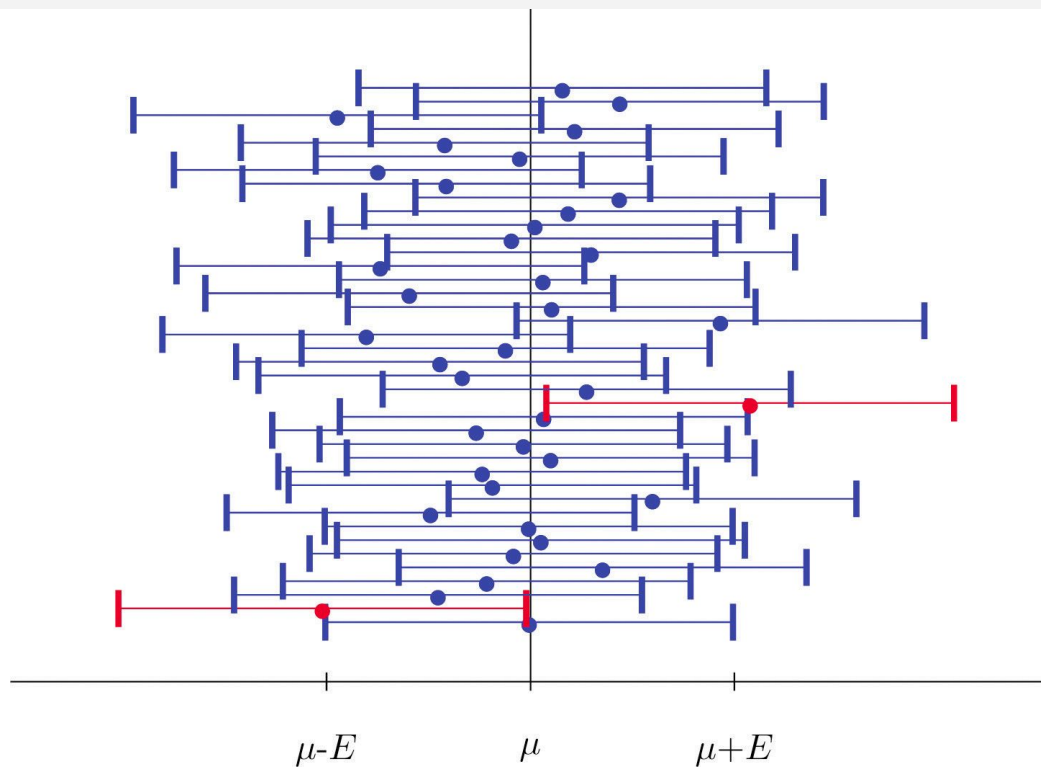
- a) 40%
- b) 60%
- c) 80%
- d) 86%
- e) 92%

INTERVALO DE CONFIANÇA

Qual o valor de $Z_{\alpha/2}$ para os seguintes níveis de confiança:

- a) 40% 0,524 =INV.NORM(0,7;0;1)
- b) 60% 0,842 =INV.NORM(0,8;0;1)
- c) 80% 1,282 =INV.NORM(0,9;0;1)
- d) 86% 1,476 =INV.NORM(0,93;0;1)
- e) 92% 1,751 =INV.NORM(0,96;0;1)

INTERVALO DE CONFIANÇA



INTERVALO DE CONFIANÇA

Suponha que um grande sindicato deseja estimar o número médio de horas por mês que um membro do sindicato está ausente do trabalho. O sindicato decide fazer uma amostragem aleatória de 25 dos seus membros e monitorizar o seu tempo de trabalho durante 1 mês. No final do mês é contabilizado o total de horas de ausência ao trabalho de cada colaborador. Se a média da amostra for 9,6 e o desvio padrão da população for 6,4 horas, encontre um intervalo de confiança de 90% para o verdadeiro número médio de horas ausentes por mês por funcionário.

- A. 9.6 ± 2.19
- B. 9.6 ± 2.64
- C. 9.6 ± 1.69
- D. 9.6 ± 2.10
- E. 9.6 ± 2.51

INTERVALO DE CONFIANÇA

Suponha que um grande sindicato deseja estimar o número médio de horas por mês que um membro do sindicato está ausente do trabalho. O sindicato decide fazer uma amostragem aleatória de 25 dos seus membros e monitorizar o seu tempo de trabalho durante 1 mês. No final do mês é contabilizado o total de horas de ausência ao trabalho de cada colaborador. Se a média da amostra for 9,6 e o desvio padrão da população for 6,4 horas, encontre um intervalo de confiança de 90% para o verdadeiro número médio de horas ausentes por mês por funcionário.

- A. 9.6 ± 2.19
- B. 9.6 ± 2.64
- C. 9.6 ± 1.69
- D. 9.6 ± 2.10
- E. 9.6 ± 2.51

$$9.6 \pm 1.645 * 6.4 / \sqrt{25}$$

INTERVALO DE CONFIANÇA

Considere que a altura média da população adulta brasileira tem distribuição normal com $\mu = 1,70$ m e $\sigma = 0,10$ m.

Se forem selecionadas várias amostras aleatórias simples de 100 pessoas,

- a) Qual é a distribuição das médias da amostra? (média e desvio padrão)
- b) Em que fração das amostras devemos obter \bar{x} entre 1,69 e 1,71?

INTERVALO DE CONFIANÇA

Considere que a altura média da população adulta brasileira tem distribuição normal com $\mu = 1,70$ m e $\sigma = 0,10$ m.

Se forem selecionadas várias amostras aleatórias simples de 100 pessoas,

a) Qual é a distribuição das médias da amostra? (média e desvio padrão)
 $1,70$ and $0,10/\text{raiz}(100) = 0,01$

b) Em que fração das amostras devemos obter x entre 1,69 e 1,71?
 $(1,69-1,70)/0,01 = -1$
 $(1,71-1,70)/0,01 = 1$

$$P(1,69 < x < 1,71) = P(-1 < Z < 1) = 0,6826$$

INTERVALO DE CONFIANÇA

$$\bar{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

INTERVALO DE CONFIANÇA

Uma pesquisa com 900 alunos foi realizada para saber como eles veem o restaurante de sua escola. A pesquisa constatou que 396 deles estavam satisfeitos. Qual é o nível de confiança de 95% do nível de satisfação?

INTERVALO DE CONFIANÇA

Uma pesquisa com 900 alunos foi realizada para saber como eles veem o restaurante de sua escola. A pesquisa constatou que 396 deles estavam satisfeitos. Qual é o nível de confiança de 95% do nível de satisfação?

$$0.44 \pm 1.96 * \sqrt{(0.44 * (1-0.44))/900)}$$

INTERVALO DE CONFIANÇA

Uma amostragem aleatória de 200 consumidores, selecionados pelos cadastros das concessionárias, indica que 82 consumidores ainda possuem os carros comprados há 2 anos. Crie intervalos de confiança para a proporção de consumidores na população que ainda possuem o carro que compraram há 2 anos com:

- A) 90%
- B) 95%
- C) 99%

INTERVALO DE CONFIANÇA

Uma amostragem aleatória de 200 consumidores, selecionados pelos cadastros das concessionárias, indica que 82 consumidores ainda possuem os carros comprados há 2 anos. Crie intervalos de confiança para a proporção de consumidores na população que ainda possuem o carro que compraram há 2 anos com:

A) 90%

$[0.41 - 1,645 \cdot 0,0348 ; 0.41 + 1,645 \cdot 0,0348]$

B) 95%

$[0.41 - 1,96 \cdot 0,0348 ; 0.41 + 1,96 \cdot 0,0348]$

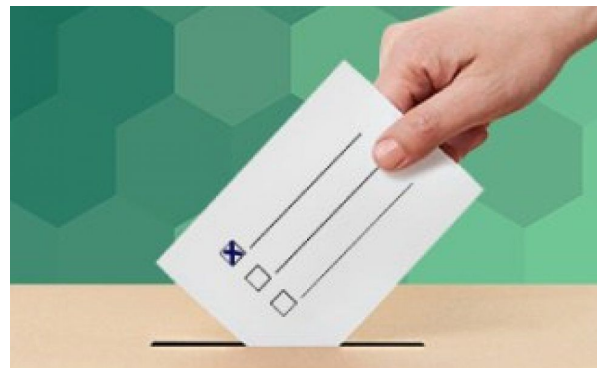
C) 99%

$[0.41 - 2,576 \cdot 0,0348 ; 0.41 + 2,576 \cdot 0,0348]$

INTERVALO DE CONFIANÇA

Para estimar a intenção de voto do candidato A em uma eleição presidencial, uma empresa entrevista 1700 pessoas.

Se 30% dizem que votarão no candidato, qual é a margem de erro dessa estimativa (com nível de confiança de 95%)?

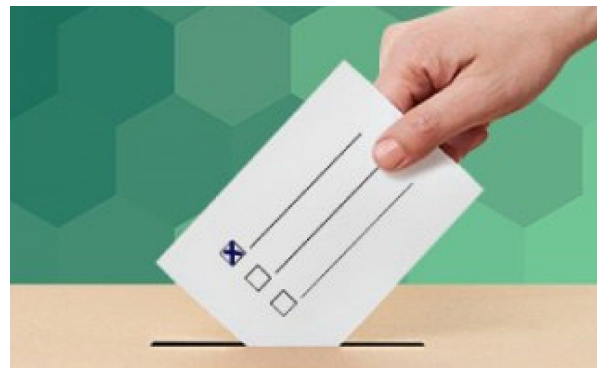


INTERVALO DE CONFIANÇA

Para estimar a intenção de voto do candidato A em uma eleição presidencial, uma empresa entrevista 1700 pessoas.

Se 30% dizem que votarão no candidato, qual é a margem de erro dessa estimativa (com nível de confiança de 95%)?

$$E = 1,96 * (0,3 * 0,7) / \text{raiz}(1700) = 1\%$$



INTERVALO DE CONFIANÇA

Calcule o intervalo de confiança de 90% para a proporção de simpatizantes do Candidato X, dado que em uma amostra aleatória simples de 200 pessoas, 75 pessoas eram simpatizantes.

- a) 32,5% to 42,5%
- b) 33,5% to 41,5%
- c) 34,1% to 40,9%
- d) 31,5% to 43,5%
- e) 32,0% to 43,0%



INTERVALO DE CONFIANÇA

Calcule o intervalo de confiança de 90% para a proporção de simpatizantes do Candidato X, dado que em uma amostra aleatória simples de 200 pessoas, 75 pessoas eram simpatizantes.

- a) 32,5% to 42,5%
- b) 33,5% to 41,5%
- c) 34,1% to 40,9%
- d) 31,5% to 43,5%
- e) 32,0% to 43,0%



Estimativa pontual $75/200 = 0.375$.

Margem de erro: $1,645 \cdot \text{raiz}(0,375 \cdot 0,625 / 200)$

TAMANHO AMOSTRAL

MARGEM DE ERRO

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Margem de Erro

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

TAMANHO AMOSTRAL

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \longrightarrow \quad n = \left(z_{\frac{\alpha}{2}} \frac{\sigma}{E} \right)^2$$

CALCULADORA

Calculadora de margem de erro

<https://pt.surveymonkey.com/mp/margin-of-error-calculator/>

EXEMPLO ELEIÇÕES

Margem de erro: 2 pontos percentuais

Nível de confiança: 95%

Proporção de votos: Distribuição Binomial =>

Variância: $p * (1-p)$

Maior variância possível:

$p = 0,5 \Rightarrow \text{Variância} = 0,25$

$$N = (1,96 * 0,5/0,02)^2 = 2,401$$



TAMANHO AMOSTRAL

Gostaríamos de estimar a renda média do motorista na plataforma Uber, com 99% de confiança, com uma precisão de R\$ 1.000. Sabemos que o desvio padrão é de R\$ 2.000. De quantos motoristas devemos obter dados?

Dica: $z(99/2)$: 2,576

TAMANHO AMOSTRAL

Gostaríamos de estimar a renda média do motorista na plataforma Uber, com 99% de confiança, com uma precisão de R\$ 1.000. Sabemos que o desvio padrão é de R\$ 2.000. De quantos motoristas devemos obter dados?

Dica: $z(99/2)$: 2,576

$$n = (2,576 * 2000)^2 / 100^2 = 2654,3$$

E SE:

- $N < 30$?

- σ é desconhecido?

DISTRIBUIÇÃO T

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \times \quad \bar{x} \pm t_{(n-1)\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

*t Distribution converges to Normal distribution as sample size increases

DISTRIBUIÇÃO T

Considere um problema onde o desvio padrão da população não é conhecido.

Quais distribuições devem ser usadas para construir um intervalo de confiança para a média: t-student ou Z (normal)?

- a) t, para qualquer tamanho de amostra
- b) t, somente se n for grande (>30)
- c) z, para qualquer tamanho de amostra
- d) z somente se n for pequeno
- e) t ou z, já que ambas as distribuições são muito semelhantes

DISTRIBUIÇÃO T

Considere um problema onde o desvio padrão da população não é conhecido.

Quais distribuições devem ser usadas para construir um intervalo de confiança para a média: t-student ou Z (normal)?

- a) t, para qualquer tamanho de amostra
- b) t, somente se n for grande (>30)
- c) z, para qualquer tamanho de amostra
- d) z somente se n for pequeno
- e) t ou z, já que ambas as distribuições são muito semelhantes

TAMANHO AMOSTRAL

$$\bar{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

Equação?

TAMANHO AMOSTRAL

$$n = \frac{(z_{\frac{\alpha}{2}})^2 p(1 - p)}{E^2}$$

TAMANHO AMOSTRAL

Qual deve ser o tamanho da amostra se com a pesquisa quisermos estimar a proporção da população com uma margem de erro de 0,025 com 95% de confiança? (com proporção de 0,44)

TAMANHO AMOSTRAL

Qual deve ser o tamanho da amostra se com a pesquisa quisermos estimar a proporção da população com uma margem de erro de 0,025 com 95% de confiança? (com proporção de 0,44)

$$n = 1.96^2 (0.44) (1 - 0.44) / (0.025)^2 = 1514.5$$

TAMANHO AMOSTRAL

Qual deve ser o tamanho da amostra se com o inquérito queremos estimar a proporção da população com uma margem de erro de 0,025 com 95% de confiança e não sabemos de antemão a proporção?

TAMANHO AMOSTRAL

Qual deve ser o tamanho da amostra se com o inquérito queremos estimar a proporção da população com uma margem de erro de 0,025 com 95% de confiança e não sabemos de antemão a proporção?

$$n = 1.96^2 (0.5) (1 - 0.5) / (0.025)^2 = 1536.6$$