



# ANÁLISE DE DADOS E DATA MINING

## MÓDULO 1

### Disciplina: Amostragem

### Tema: Amostra Aleatória Simples e Sistemática

#### Coordenação:

Professores Doutores:

Adolpho W. P. Canton

Alessandra de A. Montini

Fernando C. de Almeida

**Prof. Adolpho Walter Pimazoni Canton**

**Agosto de 2015**

**A nota da disciplina de amostragem será a nota obtida na avaliação final do curso de EAD. A avaliação final poderá ser realizada até o dia 31 de agosto de 2015.**

11	agosto	23	aula presencial
13	agosto	24	EAD
18	agosto	25	aula presencial
20	agosto	26	aula presencial

## Conteúdo

- Censo, Universo e Amostra
- Teorema do Limite Central
- Amostra Aleatória Simples
- Amostra Aleatória Sistemática
- Amostra Aleatória Estratificada
- Conceito de Margem de Erro
- Determinação do Tamanho da Amostra para inferência sobre a média para uma amostra aleatória simples
  - Considerando população finita
  - Considerando população infinita
- Determinação do Tamanho da Amostra para inferência sobre proporção para uma amostra aleatória simples
  - Considerando população finita
  - Considerando população infinita
- Determinação do Tamanho da Amostra para inferência sobre a média considerando uma amostra estratificada

**Censo**

O Censo é um levantamento de informações relacionadas a todos os elementos de uma determinada população.



De acordo com o IBGE:

A palavra censo vem do latim census e quer dizer "conjunto dos dados estatísticos dos habitantes de uma cidade, província, estado, nação".

O Censo é a única pesquisa que visita todos os domicílios brasileiros.



A palavra **Censo** vem do latim “Census” e é geralmente traduzida como levantamento, registro, estimativa.

A palavra **Censo** originalmente foi traduzida como a lista de nomes e propriedades dos cidadãos romanos.

É o particípio passado de CENSERE, “avaliar, estimar o valor de, julgar”.



Na Roma Antiga era realizado o censo para mapear os proprietários de terras e determinar o pagamento dos impostos.



**População**

# População

A população é formada por todas as observações do universo de referência.

O tamanho da população será denotado por  $N$ .

# População

## Exemplos

**Todos os aviões de um país**



**Todos os automóveis de uma cidade**



**Todos os peixes de um lago**



**Todos os brasileiros**



Algumas populações são consideradas finitas e outras populações são consideradas infinitas.

Populações finitas são aquelas que podem ser quantificadas numericamente.

Populações infinitas são aquelas que não podem ser quantificadas numericamente.

## Exemplos de população finita



Os clientes do Banco Bradesco.

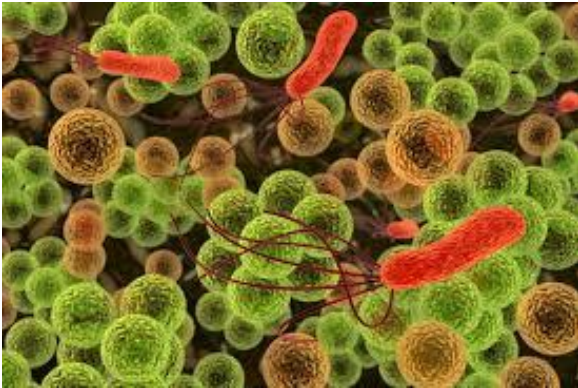


Os alunos da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo.



# Exemplos de população infinita

As Populações infinitas são geradas por um processo contínuo em que os elementos da população são gerados indefinidamente.



O número de bactérias em uma colônia de bactérias.



O número de habitantes do planeta.

**Amostra**



A amostra é formada por qualquer parte de uma população.

O tamanho da amostra será denotado por  $n$ .

# Amostra

## Exemplos

**Alguns aviões do país**



**Alguns automóveis da cidade**



**Alguns peixes do lago**



**Alguns brasileiros**



# **Teorema do Limite Central**

A média de uma população é um parâmetro populacional. Quando a média for desconhecida pode ser estimada por meio dos dados de uma amostra.

A média populacional, de uma determinada variável, será denotada por  $\mu$ .

População



# Exemplo de médias populacionais:

- Idade média
- Renda média
- Temperatura média
- Saldo médio

**População**



Quando deseja-se obter informações sobre a média da população pode-se considerar a média amostral que será denotada por  $\bar{X}$ . A média amostral é um estimador da média populacional.

**Média Populacional -  $\mu$**



**Média Amostral -  $\bar{X}$**



Cada amostra extraída de uma população gera um valor para a média amostral.

Amostra 1



**Média amostral 1**

Amostra 2



**Média amostral 2**

Amostra 3

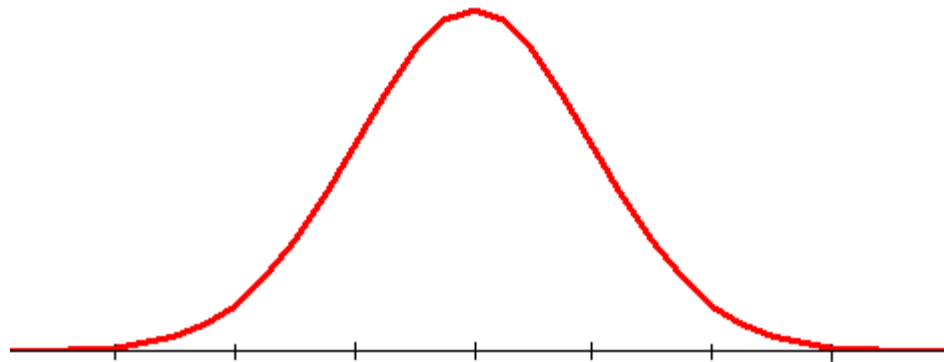


**Média amostral 3**

# Teorema do Limite Central

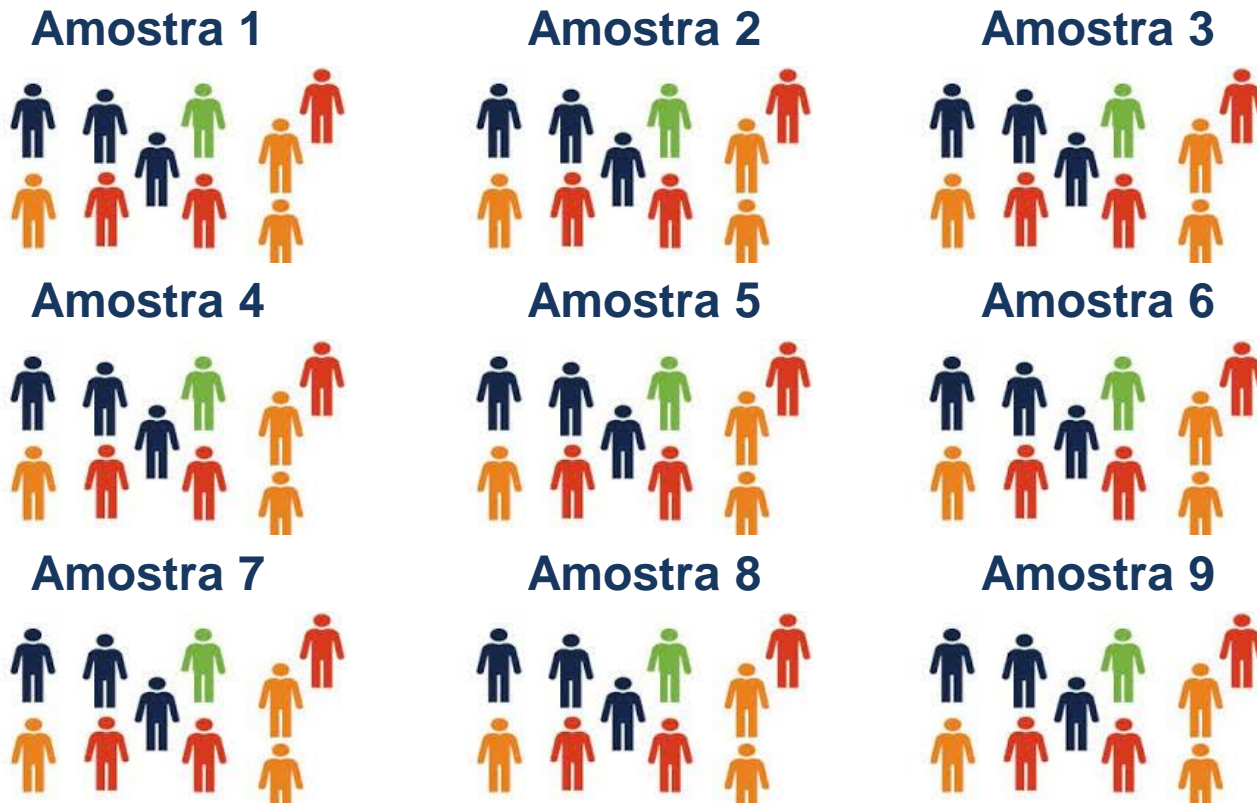
Quando a amostra possuir 30 elementos ou mais a distribuição da média amostral pode ser aproximada por uma Distribuição Normal.

Distribuição de  $\bar{X}$





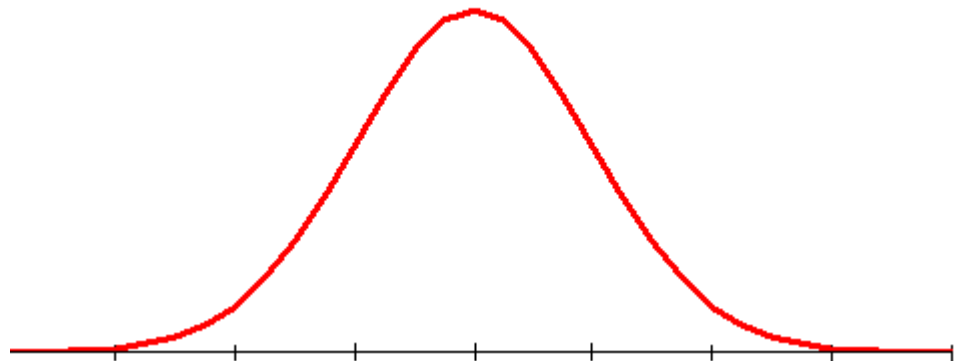
Como cada amostra extraída de uma população pode gerar um valor para a média amostral pode-se calcular o desvio padrão e a variância da média amostral. A variância da média amostral será denotada por  $\frac{\sigma^2}{n}$



# Teorema do Limite Central

A média amostral pode ser aproximada por uma Distribuição Normal com média  $\mu$  e variância  $\frac{\sigma^2}{n}$ , sendo  $\sigma^2$  a variância dos elementos.

Distribuição de  $\bar{X}$



A variância da média amostral é  $\frac{\sigma^2}{n}$  .

Como  $\sigma^2$  é a variância dos elementos da população geralmente ela é desconhecida e precisa ser estimada.

Quando a **população for infinita** o desvio padrão da média amostral é estimado por

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

em que

n é o número de elementos da amostra;

S é o desvio padrão amostral;

$$S = \sqrt{S^2}$$

$$S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$$

Quando a **população for finita** o desvio padrão da média amostral é estimado por

$$S_{\bar{x}} = \left( \frac{S}{\sqrt{n}} \right) \cdot \sqrt{\frac{N-n}{N-1}}$$

em que

N é o número de elementos da população;

n é o número de elementos da amostra;

S é o desvio padrão amostral;

# Amostragem

Será que sempre você precisa ter acesso a todos os dados de uma determinada população para realizar uma tomada de decisão relacionada à população ?

Claro que não.



Em muitos problemas do mundo real deseja-se estudar alguma variável de uma população mas por algum motivo não é possível ter acesso a toda a população. Neste caso, pode-se analisar as informações da amostra e concluir para a população.



## Exemplo 1

Quando um executivo do banco Itaú deseja saber como está a satisfação de seus clientes com relação ao atendimento do gerente de conta, ele pode ligar para toda a população de clientes ou ligar para uma amostra selecionada adequadamente.



## Exemplo 1

Neste exemplo ligar para todos os clientes gera um custo muito grande e pode levar muito tempo até conseguir contato com todos os clientes. Desta forma, a análise de uma amostra pode ser extremamente adequada.

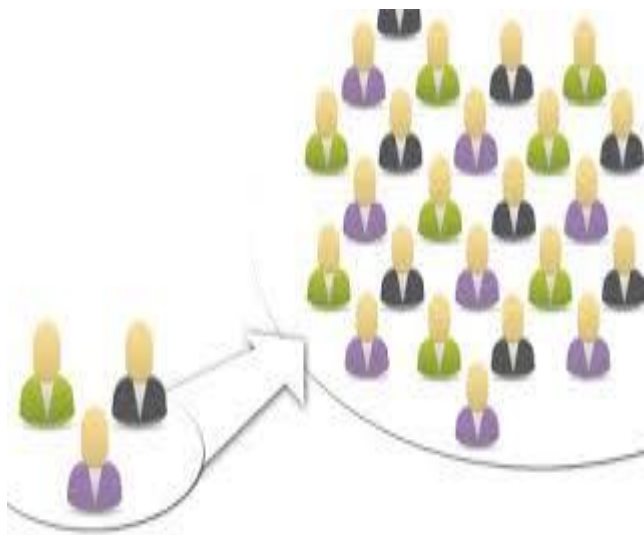


# Exemplo 1



Para resolver o problema do executivo e estimar a satisfação de seus clientes com relação ao atendimento do gerente de conta pode-se retirar uma amostra de clientes e com

base na amostra  
estimar a satisfação  
esperada para toda a  
população.



## Exemplo 2

Quando um executivo deseja saber como estão seus níveis de colesterol ele não precisa analisar todo o sangue, basta analisar o resultado de uma amostra.



Amostragem é o processo de obtenção de amostras de uma determinada população.

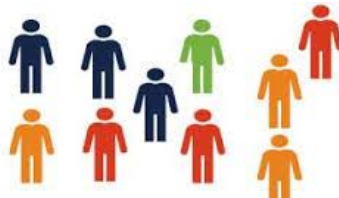
População



Amostra 1



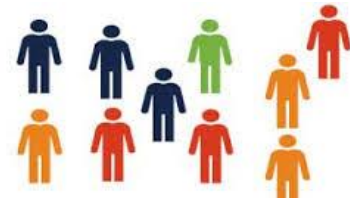
Amostra 2



.....



Amostra n





O estudo de uma amostra possibilita a obtenção de informações a respeito de **parâmetros populacionais** por meio da observação de apenas **uma amostra** da **população**.



# **Alguns métodos Probabilísticos de Amostragem**

Amostra Aleatória Simples

Amostra Aleatória Sistemática

Amostra Aleatória Estratificada



# **Amostra Aleatória Simples**

A amostragem aleatória simples é o esquema amostral mais utilizado e mais conhecido.

Este esquema amostral é fácil de ser aplicado a uma população.

Uma amostra aleatória simples é uma amostra de  $n$  elementos extraída de uma população de  $N$  elementos, em que os elementos são selecionados ao acaso.

# Exemplo 1

Como exemplo considere a seleção aleatória de  $n=1$  aluno em um grupo de  $N=11$  alunos. Neste caso todos os alunos devem possuir a mesma probabilidade de serem selecionados.



## Exemplo 2

Considere a amostra oriunda do sorteio de **uma** bola em uma urna com **N** bolas.

A bola sorteada é uma amostra aleatória simples pois todas as bolas possuem a mesma chance de serem sorteadas.



## Exemplo 3

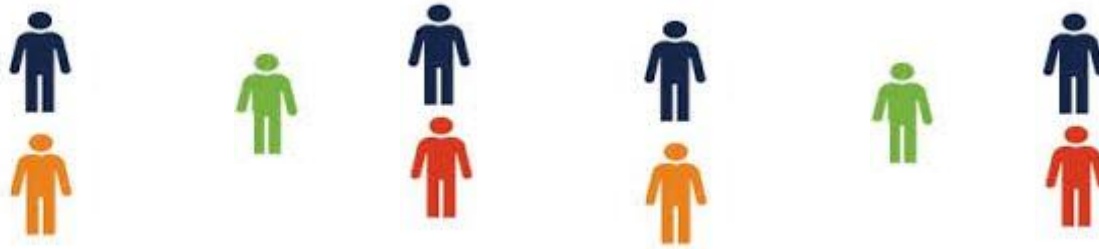
Sorteio aleatório de 1200 clientes da base de clientes do banco Itaú.

Considerando que todos os clientes possuem a mesma chance de serem sorteados esta amostra de 1200 clientes é uma amostra aleatória simples.



# Características da Amostra Aleatória Simples

- Todos os elementos da população possuem a mesma probabilidade de serem selecionados;



- Todos os subconjuntos possíveis de  $n$  elementos da população possuem a mesma probabilidade de serem selecionados.

Amostra 1



Amostra 2



Amostra 3



**Determinação do Tamanho da  
Amostra para estimação de MÉDIA  
considerando uma Amostragem  
Aleatória Simples**

# Exemplo 1

## Estimação do Ganho Médio





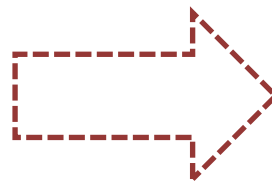


Suponha que deseja-se estimar o ganho mensal médio dos habitantes de uma determinada comunidade.

Neste caso a população é composta por todos os habitantes da comunidade que possuem ganho mensal.



Seja  **$N=15.000$**  o número de habitantes da comunidade que possuem ganho mensal.



Seja  $\mu$  o ganho mensal médio dos habitantes da comunidade.

Como  $\mu$  é um valor relacionado a toda a população ele é denominado de **parâmetro populacional**.

**Parâmetro** é a descrição numérica de uma determinada característica de uma **população**.

No exemplo o valor de  $\mu$  não é conhecido e precisa ser estimado.

**Estratégia:** Retira-se uma amostra de tamanho  $n$  da população e estima-se o valor de  $\mu$ .

Para estimar o valor de  $\mu$  deve ser calculada a média amostral, denominada  $\bar{X}$

Média Amostral =  $\bar{X}$



Média Populacional =  $\mu$

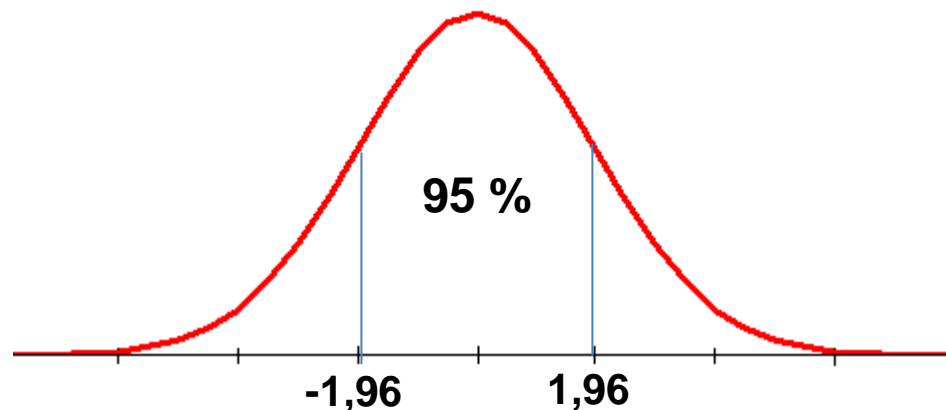


Para estimar a média amostral é necessário a extração de uma amostra aleatória simples de tamanho  $n$ .



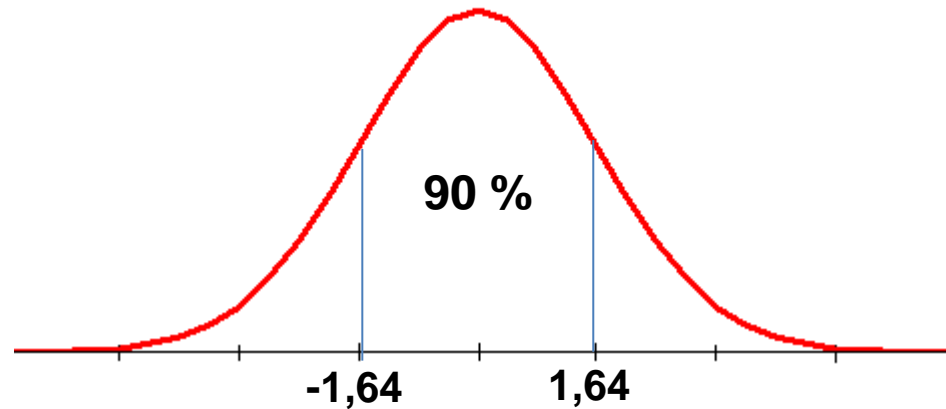
# Qual o valor de $n$ ?

Para a obtenção do valor de  $n$  é necessário definir a confiança a ser utilizada. Geralmente considera-se 95 % de confiança. A média amostral possui distribuição Normal quando são selecionados pelo menos 30 elementos. Desta forma, o valor que gera uma confiança de 95 % será obtido por meio da Distribuição Normal Padrão ( $Z=1,96$ ).



# Qual o valor de $n$ ?

Caso o objetivo seja considerar 90 % de confiança o valor obtido por meio da distribuição Normal Padrão é  $Z=1,64$ .



# Qual o valor de $n$ ?

Para a obtenção do valor de  $n$  é necessário definir a margem de erro (ME) do estudo.

A margem de erro (ME) é a diferença esperada entre o valor do parâmetro populacional  $\mu$  e o valor obtido com a média amostral  $\bar{X}$ .

# Qual o valor de $n$ ?

## Exemplo de margem de erro

Considere a afirmação:

Uma pesquisa indica que o ganho mensal médio dos habitantes de uma comunidade é **R\$ 1200,00** com uma margem de erro de **R\$ 200,00**, para uma confiança de 95 %.

Esta afirmação quer dizer que com **95 % de probabilidade** o verdadeiro ganho mensal médio dos habitantes da comunidade  $\mu$  pode ser um valor entre **R\$ 1.000,00 e R\$ 1.400,00**.



Como a população da comunidade é finita e pretende-se estimar a média populacional o tamanho da amostra a ser retirada é dado pela expressão.

$$n = \frac{Z^2 \cdot S^2 \cdot N}{(ME)^2 \cdot (N - 1) + Z^2 \cdot S^2}$$

Em que :

**ME** é a margem de erro do estudo;

**Z** é o valor da Distribuição Normal que fornece a confiança desejada;

**N** é o tamanho da população;

**S<sup>2</sup>** é a variância estimada em uma amostra piloto.

No exemplo, deseja-se estimar o ganho mensal médio dos habitantes de uma determinada comunidade.

Seja  **$N=15.000$**  o número de habitantes da comunidade que possuem ganho mensal.

Considere que a margem de erro (**ME**) é R\$ **100,00**.

Considere 95 % de confiança. Desta forma, o valor obtido por meio da Distribuição Normal Padrão é  **$Z=1,96$** .

Inicialmente deve-se extrair uma amostra piloto para obter o valor de  $S^2$ .

Suponha que neste exemplo, foi extraída uma amostra piloto e originou um desvio padrão  $S=500$ . Desta forma,  $S^2=250.000$ .

O tamanho da amostra é dado por:

$$n = \frac{Z^2 \cdot S^2 \cdot N}{(ME)^2 \cdot (N - 1) + Z^2 \cdot S^2}$$

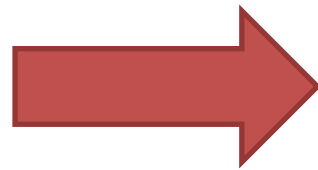
$$n = \frac{(1,96)^2 \cdot (250.000) \cdot (15.000)}{(100)^2 \cdot (15000 - 1) + (1,96)^2 \cdot (250.000)} = 95$$

# Qual o valor de $n$ ?

Para estimar o ganho mensal médio dos habitantes de uma determinada comunidade, considerando uma amostra aleatória simples, deve-se extrair uma amostra de **95 elementos**.



Considere que retirou-se uma amostra aleatória simples de tamanho  $n = 95$  elementos da comunidade e obteve-se uma média amostral de R\$ 1.200,00.



$$\bar{X} = 1.200,00$$

O valor **1.200,00** é a estimativa pontual do verdadeiro ganho mensal médio dos habitantes da comunidade.

A estimativa por intervalo de uma média populacional é dada pelo intervalo de confiança para  $\mu$  dado por:

$$\bar{X} \pm \text{margem de erro}$$

No exemplo, a estimativa por intervalo do verdadeiro para o ganho mensal médio dos habitantes da comunidade, considerando 95 % de confiança é dada por:

$$\overline{X} \mp \text{margem de erro}$$

$$\text{R\$ 1.200,00} \pm \text{R\$ 100,00}$$



Desta forma, o ganho mensal médio dos habitantes da comunidade pode estar entre **R\$ 1.100,00 e R\$ 1.300,00** com 95 % de confiança.

O intervalo de confiança para  $\mu$  dado por:

$$\bar{X} \mp \text{margem de erro}$$

A margem de erro, para uma população finita é dada por:

$$Z \cdot \left( \frac{S}{\sqrt{n}} \right) \cdot \sqrt{\frac{N-n}{N-1}}$$



**Determinação do Tamanho da  
Amostra, para estimação de MÉDIA ,  
para População Finita considerando  
um amostra aleatória simples**

Como a população da comunidade é finita e pretende-se estimar a média populacional o tamanho da amostra a ser retirada é dado pela expressão.

$$n = \frac{Z^2 \cdot S^2 \cdot N}{(ME)^2 \cdot (N - 1) + Z^2 \cdot S^2}$$

Em que :

**ME** é a margem de erro do estudo

**Z** é o valor da Distribuição Normal que fornece a confiança desejada

**N** é o tamanho da população

**S<sup>2</sup>** é a variância estimada em uma amostra piloto.

**Determinação do Tamanho da  
Amostra, para estimação de MÉDIA ,  
para População Infinita considerando  
um amostra aleatória simples**

Quando a população é muito grande (infinita) e pretende-se estimar a média populacional o tamanho da amostra a ser retirada é dado pela expressão.

$$n = \frac{Z^2 \cdot S^2}{(ME)^2}$$

Em que :

**ME** é a margem de erro do estudo

**Z** é o valor da Distribuição Normal que fornece a confiança desejada

**S<sup>2</sup>** é a variância estimada por uma amostra piloto.

O intervalo de confiança para  $\mu$  dado por:

$$\bar{X} \pm \text{margem de erro}$$

A margem de erro, para uma população infinita é dada por:

$$Z \cdot \left( \frac{s}{\sqrt{n}} \right)$$

## Exemplo 2

# Estimação da Proporção de Eleitores





Suponha que deseja-se estimar a proporção de eleitores que possuem intenção em votar na candidata Yeda.

Neste caso a população é composta por todos os habitantes da comunidade que possuem direito a voto.



Seja  **$N=8.000$**  o número de habitantes da comunidade com direito a voto.



Seja  $\pi$  a proporção populacional de eleitores que possuem intenção em votar na candidata Yeda.

Como  $\pi$  é um valor relacionado a toda a população ele é denominado **parâmetro populacional**.

**Parâmetro** é a descrição numérica de uma determinada característica de uma população.



No exemplo, o valor de  $\pi$  não é conhecido e precisa ser estimado.

**Estratégia:** Retira-se uma amostra de tamanho  $n$  da população para estimar o valor de  $\pi$ .

Para estimar o valor de  $\pi$  deve ser calculada a proporção amostral de pessoas com intenção em votar na candidata Yeda, denominada  $p$ .

Proporção Amostral =  $p$

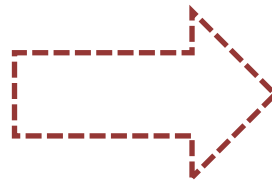


Proporção Populacional =  $\pi$



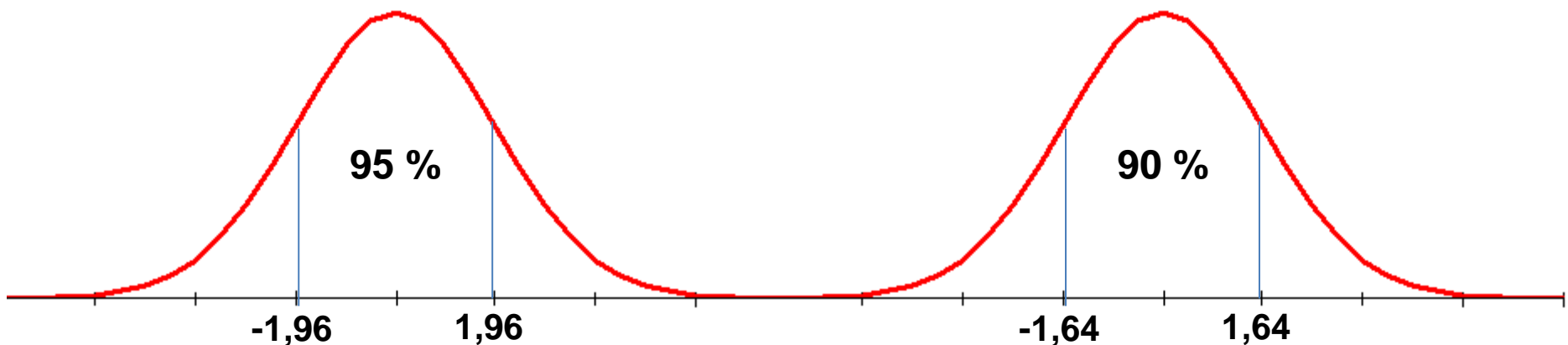
Para estimar a proporção amostral de pessoas com intenção em votar na candidata Yeda, é necessário a extração de uma amostra aleatória simples de tamanho  $n$ .

Amostra de tamanho  $n$



# Qual o valor de $n$ ?

Para a obtenção do valor de  $n$  é necessário definir a confiança a ser utilizada. A proporção amostral possui Distribuição Normal quando são selecionados pelo menos 30 elementos. Desta forma, os valores que geram confianças de 90% e 95 % serão obtidos por meio da distribuição Normal Padrão e são respectivamente  **$Z=1,96$**  e  **$Z=1,64$** .



# Qual o valor de $n$ ?

Para a obtenção do valor de  $n$  é necessário definir a margem de erro (ME) do estudo.

A margem de erro (ME) é a diferença esperada entre o valor do parâmetro populacional  $\pi$  e o valor estimado do parâmetro  $p$  obtido pela amostra.

# Qual o valor de $n$ ?

## Exemplo de margem de erro

Considere a afirmação:

Uma pesquisa indica que a proporção amostral de pessoas com intenção de votar na candidata Yeda é **0,75** com uma margem de erro de **0,03**, para uma confiança de 95 %.

Esta afirmação quer dizer que com **95 % de probabilidade** a proporção amostral de pessoas com intenção de votar na candidata Yeda pode ser um valor entre **0,72 e 0,78**.

Como a população da comunidade é finita e pretende-se estimar a proporção populacional de pessoas com intenção em votar na candidata Yeda, considerando a margem de erro máxima, o tamanho da amostra a ser retirada é dado pela expressão.

$$n = \frac{Z^2 \cdot 0,25 \cdot N}{(ME)^2 \cdot (N - 1) + Z^2 \cdot 0,25}$$

Em que :

**ME** é a margem de erro do estudo;

**N** é o tamanho da população;

**Z** é o valor da Distribuição Normal que fornece a Confiança desejada.

No exemplo, deseja-se estimar a proporção de eleitores que possuem intenção em votar na candidata Yeda.



Seja  **$N=8.000$**  o número de habitantes da comunidade com direito a voto.

Considere que a margem de erro (**ME**) é **0,03**.

Considere 95 % de confiança. Desta forma, o valor obtido por meio da Distribuição Normal Padrão é  **$Z=1,96$** .



Para estimar a proporção amostral de pessoas com intenção em votar na candidata Yeda, por meio de uma amostra aleatória simples, o tamanho da amostra é dado por:

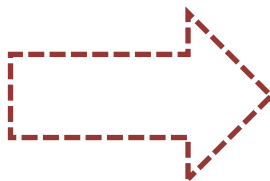
$$n = \frac{Z^2 \cdot 0,25 \cdot N}{(ME)^2 \cdot (N - 1) + Z^2 \cdot 0,25}$$

$$n = \frac{(1,96)^2 \cdot 0,25 \cdot (8.000)}{(0,03)^2 \cdot (8.000 - 1) + (1,96)^2 \cdot 0,25} = 941,62 = 942$$



Considere que retirou-se uma amostra aleatória simples de tamanho  **$n = 942$**  da comunidade.

Considere que na amostra a proporção amostral de pessoas com intenção em votar na candidata Yeda é



$$p = 0,65$$

O valor  **$0,65$**  é a estimativa pontual da verdadeira proporção de pessoas com intenção de votas na candidata Yeda.

A estimativa por intervalo para uma proporção populacional é dada pelo intervalo de confiança para a proporção populacional que é dado por:

$$p \pm \text{margem de erro}$$

A margem de erro, para uma população finita é dada por:

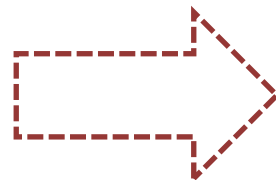
$$Z \cdot \left( \sqrt{\frac{0,25}{n}} \right) \cdot \sqrt{\frac{N - n}{N - 1}}$$

No exemplo, a estimativa por intervalo para a verdadeira proporção de pessoas com intenção em votar na candidata Yeda, considerando 95 % de confiança é dada por:

$$p \pm \text{margem de erro}$$

$$0,65 \pm 0,03$$

Desta forma, a proporção populacional de pessoas com intenção em votar na candidata Yeda pode estar entre **0,62 e 0,68** com 95 % de confiança.



**0,62 e 0,68**

Quando a população é muito grande (infinita) e pretende-se estimar uma proporção, considerando a margem de erro máxima, o tamanho da amostra a ser retirada é dado pela expressão.

$$n = \frac{Z^2 . 0,25}{(ME)^2}$$

Em que :

**ME** é a margem de erro do estudo;

**Z** é o valor da Distribuição Normal que fornece a confiança desejada.

**Tamanho da Amostra, para estimação  
de proporção, para População Finita,  
considerando uma amostra aleatória  
simples**

Como a população da comunidade é finita e pretende-se estimar a proporção populacional de pessoas com intenção em votar na candidata Yeda média, considerando a margem de erro máxima, o tamanho da amostra a ser retirada é dado pela expressão.

$$n = \frac{Z^2 \cdot 0,25 \cdot N}{(ME)^2 \cdot (N - 1) + Z^2 \cdot 0,25}$$

Em que :

**ME** é a margem de erro do estudo;

**N** é o tamanho da população;

**Z** é o valor da Distribuição Normal que fornece a Confiança desejada.

A estimativa por intervalo para uma proporção populacional é dada pelo intervalo de confiança para a proporção populacional que é dado por:

$$p \pm \text{margem de erro}$$

A margem de erro, para uma população finita é dada por:

$$Z \cdot \left( \sqrt{\frac{0,25}{n}} \right) \cdot \sqrt{\frac{N - n}{N - 1}}$$

**Tamanho da Amostra, para estimação  
de proporção, para População  
INFINITA, considerando uma amostra  
aleatória simples**



Quando a população é muito grande (infinita) e pretende-se estimar uma proporção, considerando a margem de erro máxima, o tamanho da amostra a ser retirada é dado pela expressão.

$$n = \frac{Z^2 . 0,25}{(ME)^2}$$

Em que :

**ME** é a margem de erro do estudo;

**Z** é o valor da Distribuição Normal que fornece a Confiança desejada;

A estimativa por intervalo para uma proporção populacional é dada pelo intervalo de confiança para a proporção populacional que é dado por:

$$p \pm \text{margem de erro}$$

A margem de erro, para uma população infinita é dada por:

$$Z \cdot \left( \sqrt{\frac{0,25}{n}} \right)$$

# EXERCÍCIO

Banco de Dados : Dados\_credito.xls

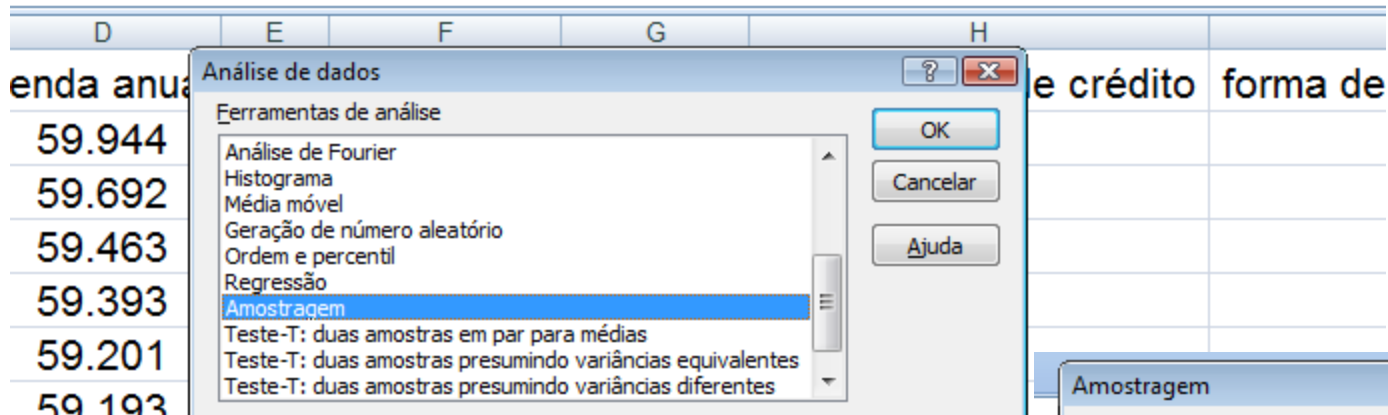
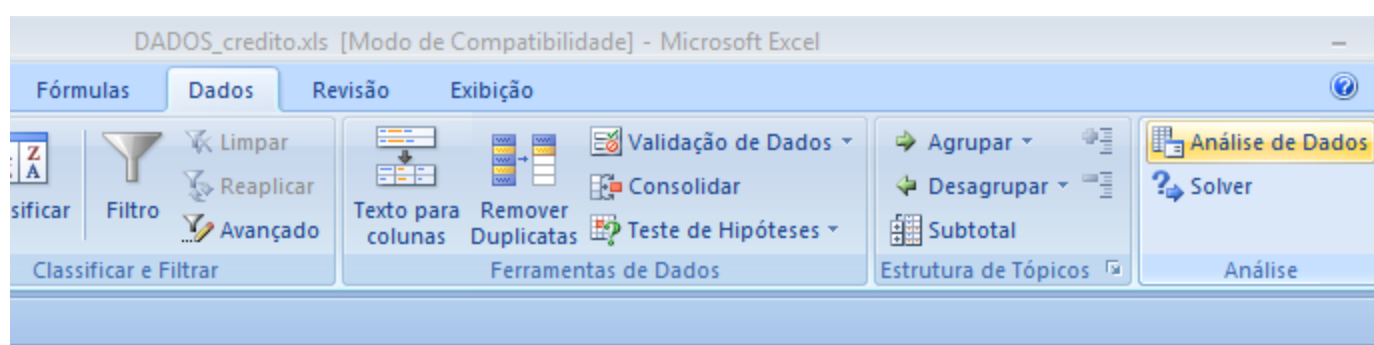
N = 2455 clientes

sexo	1	F
	2	M
estado civil	1	divorciado
	2	casado
	3	solteiro
forma de pagamento	1	mensal
	2	semanal
crédito imobiliário	1	não
	2	sim
risco de credito	1	alto
	2	medio
	3	baixo

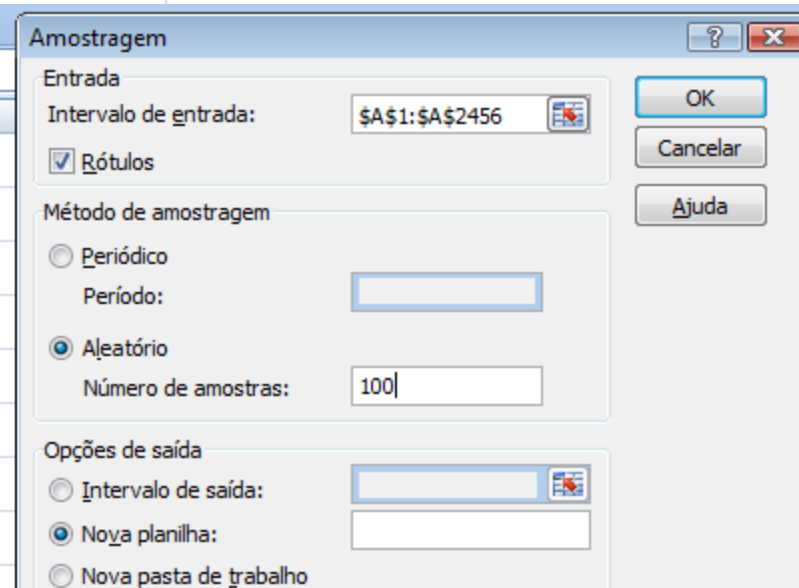
- idade (anos)
- renda anual (reais)
- número de filhos
- número de cartões de crédito
- número de cartões
- outros empréstimos (quantidade)

## EXERCÍCIO

- 1) Obter a média populacional e o desvio padrão populacional para as variáveis: idade e renda anual;
- 2) Obter a distribuição de frequência para a variável sexo;
- 3) Obter o tamanho da amostra para o estudo da variável renda de uma população finita, considerando uma margem de erro de R\$ 1.670,00, 95 % de confiança e um desvio padrão populacional estimado de R\$ 8.700,00;
- 4) Obter uma amostra aleatória simples sem reposição por meio do excel para o tamanho da amostra obtido no exercício 3.



**Amostra Aleatória Simples Sem reposição de tamanho  $n = 100$**



Colar a amostra gerada no exercicio 4 na planilha: amostra\_alunos\_n=100

	A	B	C	D
1	310			
2	1015			
3	1126			
4	1947			
5	296			
6	598			
7	2339			
8	314			
9	565			
10	1849			
11	665			
12	829			
13	243			
14	2002			
15	1481			
16	661			
17	1749			
18	1144			
19	1144			
20	598			
21	1734			
22	732			
23	1840			
24	1947			
25	598			
26	1840			
27	243			
28	554			
29	17			

	A	B	C	
1	n	identificaca	idade	r
2	310	100092	32	
3	1015	100935	18	
4	1126	101951	25	
5	1947	100808	18	
6	296	100119	34	
7	598	103855	43	
8	2339	102098	20	
9	314	100103	33	
10	565	103418	46	
11	1849	102335	22	
12	665	103801	41	
13	829	103556	27	
14	243	100719	31	
15	2002	103120	43	
16	1481	101145	20	

- 5) Obter um intervalo com 95 % de confiança para a renda média populacional para a amostra gerada no exercício 4.
- 6) Qual o tamanho da amostra necessário para o estudo da variável renda de uma população finita, considerando uma margem de erro de R\$ 1.000,00, 95 % de confiança e um desvio padrão populacional estimado pelo desvio padrão da amostra gerada no exercício 4.
- 7) Obter um intervalo com 95 % de confiança para a proporção de pessoas do sexo feminino considerando a amostra gerada no exercício 4.
- 8) Obter o tamanho da amostra para o estudo da proporção de pessoas que possuem crédito imobiliário e uma população finita ( $N = 2455$ ), considerando uma margem de erro de 0,03 e 95 % de confiança. Considerar  $\hat{P} = 0,5$ .
- 9) Obter uma amostra aleatória simples sem reposição por meio do excel para o tamanho da amostra obtido no exercício 8.
- 10) Obter um intervalo com 95 % de confiança para a proporção de pessoas com crédito imobiliário na amostra gerada no exercício 9.

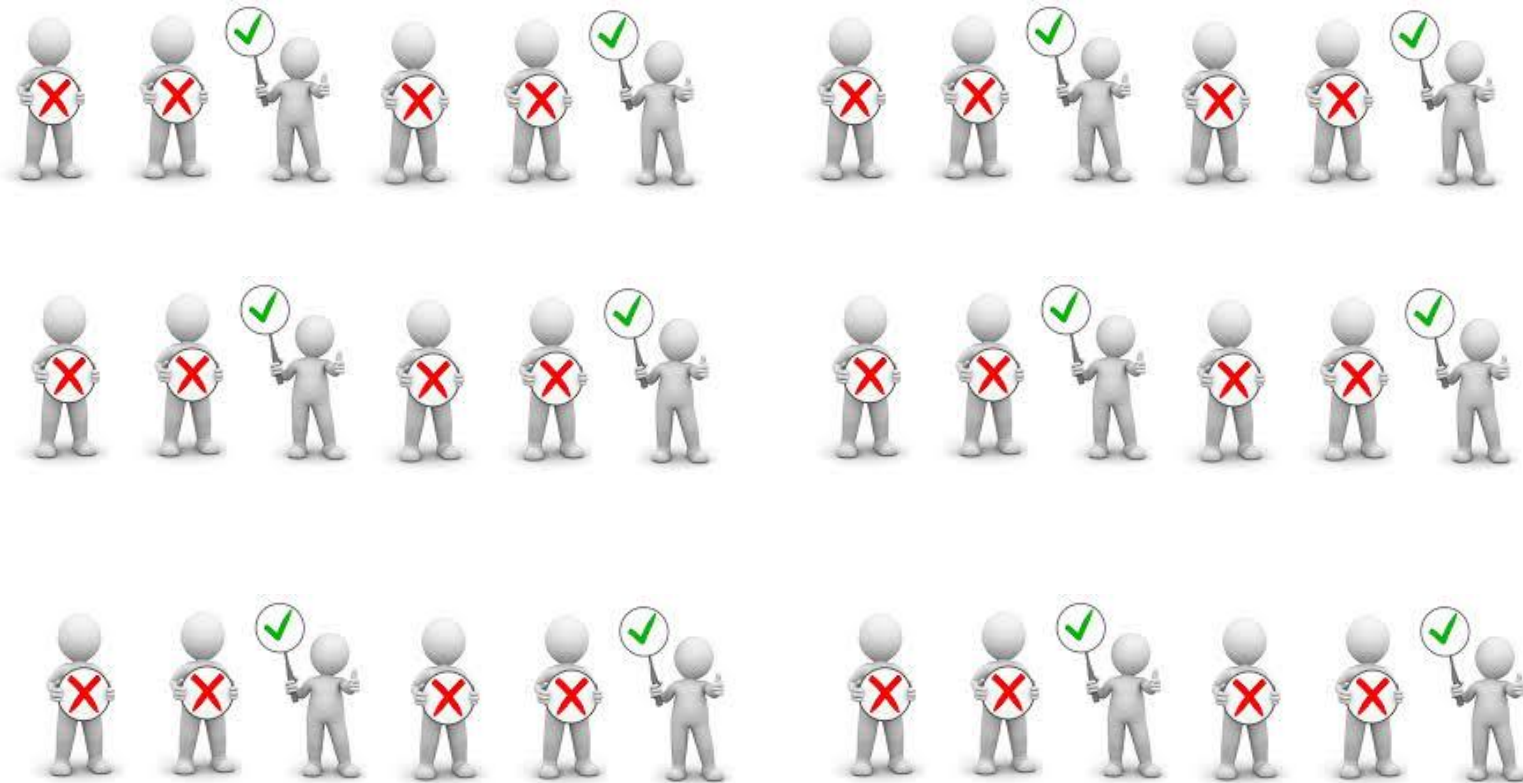
# **Amostra Aleatória Sistemática**



A amostragem aleatória sistemática é um esquema amostral em que os elementos são selecionados a partir de um critério que é aplicado a toda a população de forma sistemática.



Pode-se por exemplo selecionar um elemento a cada dois elementos da população.



Quando se deseja extrair uma amostra sistemática de tamanho  $n$  de uma população de tamanho  $N$ , deve-se amostrar um elemento a cada  $\frac{N}{n}$  elementos da população.

## Exemplo:

Deseja-se obter uma amostra sistemática de **n=2.000** clientes de uma base de dados de **N=12.000** clientes de uma empresa.

Como  $\frac{12000}{2000} = 6$  deve-se selecionar sistematicamente

de 6 em 6 clientes.

# Exemplo:

## Procedimento:

1. Deve-se ordenar a base de dados populacional por algum critério. Por exemplo ordenar a base por idade, renda, tempo de relacionamento.

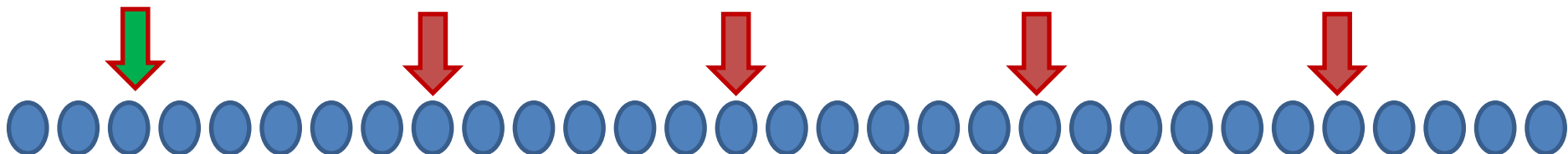
2. Deve-se selecionar aleatoriamente um elemento dos primeiros  $\frac{N}{n}$  elementos da lista populacional.

Suponha que entre os 6 primeiros elementos o elemento selecionado foi o terceiro.

3. De forma sistemática deve-se selecionar os elementos de 6 em 6 a partir do terceiro elemento.

### Sorteado

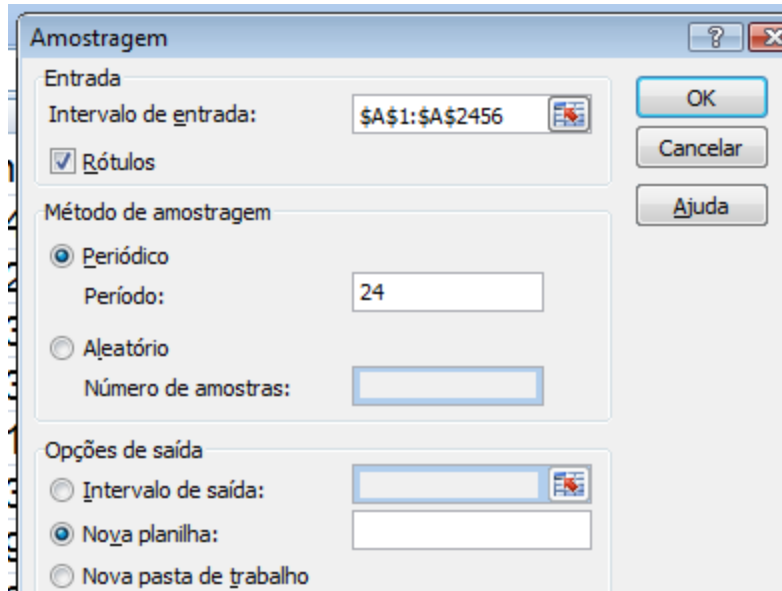
Aleatoriamente  
entre os 6  
primeiros



# EXERCÍCIO

1) Obter uma amostra sistemática de tamanho 100, por meio do excel. Considerar o banco de dados : Dados\_credito.xls

N = 2455 clientes



$$\frac{2455}{100} = 24,5$$

# **Amostra Aleatória Estratificada**



Deve-se extrair uma amostra estratificada quando a população é composta por estratos.





# Exemplo

## Gasto Médio de Famílias com Cartão de Crédito em um Supermercado



Suponha que deseja-se estimar o gasto médio com cartão de crédito dos clientes em um supermercado.



Suponha que a população é composta por três estratos. Considere o Estrato 1 os usuários do **Mastercard**, o Estrato 2 os usuários do **Visa** e o Estrato 3 os usuários do **American Express**.

Suponha que a população é composta por:



### **Estrato 1**

$N_1 = 8.000$  usuários do cartão **Mastercard**



### **Estrato 2**




$N_2 = 7.000$  usuários do cartão **Visa**



### **Estrato 3**

$N_3 = 2.000$  usuários do American **Express**

A população de clientes é distribuída da seguinte forma:

Estrato		N	%	W
Estrato 1		8000	47,06	0,47 → $W_1$
Estrato 2		7000	41,18	0,41 → $W_2$
Estrato 3		2000	11,76	0,12 → $W_3$
Total		17000		

Para estimar o gasto médio com cartão de crédito dos clientes no supermercado suponha que o gerente de Marketing retirou uma amostra estratificada de  **$n=1.500$**  clientes distribuída da seguinte forma



### **Amostra do Estrato 1**

**$n_1=705$  usuários do cartão Mastercard**



### **Amostra do Estrato 2**

**$n_2=617$  usuários do cartão Visa**



### **Amostra do Estrato 3**

**$n_3=178$  usuários do American Express**

Com base na amostra obteve-se o valor gasto médio:



## Amostra do Estrato 1 - $n_1=705$

Gasto médio =  $\bar{X}_1 = \text{R\$ } 650,00$



## Amostra do Estrato 2 - $n_2=617$

Gasto médio =  $\bar{X}_2 = \text{R\$ } 450,00$



## Amostra do Estrato 3 - $n_3=178$

Gasto médio =  $\bar{X}_3 = \text{R\$ } 1250,00$

O estimador da média populacional, calculado por meio da amostragem estratificada, é dado por:

$$\bar{X}_{st} = w_1 \bar{X}_1 + w_2 \bar{X}_2 + w_3 \bar{X}_3$$

A Tabela apresenta o peso de cada estrato e a média amostral do gasto médio do cliente em cada estrato.

Estrato	W	Média Amostral
Estrato 1	0,47	R\$ 650,00
Estrato 2	0,41	R\$ 450,00
Estrato 3	0,12	R\$ 1.250,00

A média populacional estimada é dado por:

$$\overline{X}_{st} = 0,47 * 650 + 0,41 * 450 + 0,12 * 1250 = 640,00$$





Neste exemplo pode-se concluir que em média os clientes gastam com o cartão de crédito R\$ 650,00.





Alguns clientes gastam no cartão de crédito mais do que R\$ 650,00 e outros gastam menos do que R\$ 650,00.

Qual o desvio padrão do gasto em cartão de crédito ?

Para cada estrato pode-se obter a variância e o desvio padrão do gasto do cliente.



## Amostra do Estrato 1

$$\text{Variância} = S_1^2 = \text{R\$ } 2.500,00$$

$$\text{Desvio padrão} = S_1 = \text{R\$ } 50,00$$



## Amostra do Estrato 2

$$\text{Variância} = S_2^2 = \text{R\$ } 6.400,00$$

$$\text{Desvio padrão} = S_2 = \text{R\$ } 80,00$$



## Amostra do Estrato 3

$$\text{Variância} = S_3^2 = \text{R\$ } 12.100,00$$

$$\text{Desvio padrão} = S_3 = \text{R\$ } 110,00$$

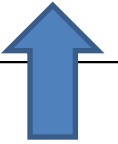


A variância de  $\bar{X}_{st}$  é dada por:

$$V(\bar{X}_{st}) = (W_1)^2 \left( \frac{S_1^2}{n_1} \right) \cdot \frac{(N_1 - n_1)}{N_1 - 1} + (W_2)^2 \left( \frac{S_2^2}{n_2} \right) \cdot \frac{(N_2 - n_2)}{N_2 - 1} + (W_3)^2 \left( \frac{S_3^2}{n_3} \right) \cdot \frac{(N_3 - n_3)}{N_3 - 1}$$

O desvio padrão é obtido por meio da raiz quadrada da variância.

$$(W_h)^2 \left( \frac{S_h^2}{n_h} \right) \cdot \frac{(N_h - n_h)}{N_h - 1}$$



Estrato	Nh	W	$(W)^2$	nh	Variância	$\left( \frac{S_h^2}{n_h} \right)$	$\frac{(N_h - n_h)}{N_h - 1}$	
Estrato 1	8000	0,47	0,22	282	R\$ 2.500,00	8,87	0,9649	1,89
Estrato 2	7000	0,41	0,17	246	R\$ 6.400,00	26,02	0,9650	4,22
Estrato 3	2000	0,12	0,01	72	R\$ 12.100,00	168,06	0,9645	2,33

A variância de  $\bar{X}_{st}$  é dada por:

$$1,89 + 4,22 + 2,33 = 8,44$$

A variância de  $\bar{X}_{st} = 8,44$

O desvio padrão de  $\bar{X}_{st} = \sqrt{8,44} = 2,905$

A estimativa por intervalo, com 95 % de confiança, para a média populacional é dada por:

$$\bar{X}_{st} \pm 1,96 * (\text{desvio padrão de } \bar{X}_{st} )$$

A estimativa por intervalo, com 95 % de confiança, para o gasto médio dos clientes é dada por:

$$\bar{X}_{st} \pm 1,96 * (\text{desvio padrão de } \bar{X}_{st} )$$

$$650 \pm 1,96 * 2,905$$

$$650 \pm 5,69$$

Com 95 % de confiança, para o verdadeiro gasto médio dos clientes está entre 644,31 e 655,69.



**Tamanho da Amostra, para estimação  
de MÉDIA , para População Finita  
considerando um amostra aleatória  
estratificada**

O tamanho da amostra a ser retirada quando a população é finita, considerando três estratos, é dado pela expressão.

$$n = \frac{N(N_1.S_1^2 + N_2.S_2^2 + N_3.S_3^2)}{N^2.D^2 + (N_1.S_1^2 + N_2.S_2^2 + N_3.S_3^2)}$$

Em que :

$$D^2 = \frac{d^2}{Z^2} \quad , \quad d \text{ é a precisão desejada}$$

**Z** é o valor da Distribuição Normal que fornece a confiança desejada;

**N** é o tamanho da população e  $N_1$ ,  $N_2$ ,  $N_3$  é o tamanho dos estratos na população;

$S_1^2$   $S_2^2$   $S_3^2$  são as variâncias dos estratos a serem estimadas por uma amostra piloto.

O tamanho da amostra a ser retirada em cada estrato, quando a população é finita, é dado pela expressão.

$$n_h = \frac{N_h}{N} . n$$

Em que :

**N** é o tamanho da população;

**n** é o tamanho da amostra;

**N<sub>1</sub>, N<sub>2</sub>, N<sub>3</sub>** é o tamanho dos estratos na população;

# **Determinação do Tamanho da Amostra para Variável Quantitativa**

## **Alguns métodos para a locação da amostra:**

- Igual tamanho de amostra para cada estrato;
- Alocação proporcional;
- Alocação de Neyman - custos amostrais iguais entre os estratos;

O tamanho da amostra  $n_h$  para esses métodos são:

Igual tamanho de amostra para cada estrato  $n_h = \frac{n}{L}$

Alocação proporcional  $n_h = \frac{N_h}{N} \cdot n$

Alocação de Neyman  $n_h = \frac{N_h \cdot S_h}{\sum_h (N_h \cdot S_h)} \cdot n$

O tamanho total n da amostra é determinado por meio da expressão :

$$d^2 = Z^2 . V(\bar{x}_{st})$$

$$V(\bar{x}_{st}) = \frac{d^2}{Z^2} = D^2$$

Em que:

d é a precisão desejada;

Z é obtido por meio da confiança definida, em geral = 95 %, ou seja, Z=1,96;

A amostra total,  $n$ , em cada um dos 4 casos é dado por:

Igual tamanho de amostra para cada estrato 
$$n = \frac{L \sum N_h^2 \cdot S_h^2}{N^2 \cdot D^2 + \sum N_h \cdot S_h^2}$$

Alocação proporcional 
$$n = \frac{N \sum N_h \cdot S_h^2}{N^2 \cdot D^2 + \sum N_h \cdot S_h^2}$$

Alocação de Neyman 
$$n = \frac{(\sum N_h \cdot S_h)^2}{N^2 \cdot D^2 + \sum N_h \cdot S_h^2}$$



# **Determinação do Tamanho da Amostra para Proporção**

## **Alguns métodos para a locação da amostra:**

- Alocação proporcional;
- Alocação de Neyman - custos amostrais iguais entre os estratos;

O tamanho da amostra  $n_h$  para esses métodos são:

Alocação proporcional  $n_h = \frac{N_h}{N} \cdot n$

Alocação de Neyman  $n_h = \frac{N_h \cdot \sqrt{P_h \cdot (1 - P_h)}}{\sum_h (N_h \cdot \sqrt{P_h \cdot (1 - P_h)})} \cdot n$

O n total em cada um dos 4 casos é dado por:

$$d = |p_{st} - P| \quad D = \frac{d}{Z}$$

Alocação proporcional

$$n = \frac{N \cdot \sum N_h \cdot P_h \cdot (1 - P_h)}{N^2 \cdot D^2 + \sum N_h \cdot P_h \cdot (1 - P_h)}$$

Alocação de Neyman

$$n = \frac{\left( \sum N_h \cdot \sqrt{P_h \cdot (1 - P_h)} \right)^2}{N^2 \cdot D^2 + \sum_h N_h \cdot P_h \cdot (1 - P_h)}$$

## Exercício

1) Extrair uma amostra aleatória simples em cada estrato do banco de dados:

### **benfeitores.xls**

O banco de dados possui: 13.637 casos

Variável relacionada ao estrato : spendest

Spendest = 1 retirar uma amostra de tamanho 30

Spendest = 2 retirar uma amostra de tamanho 38

Spendest = 3 retirar uma amostra de tamanho 18

Spendest = 4 retirar uma amostra de tamanho 143

2) Obter a média e a variância da variável `expend_tot` para cada estrato.

3) Obter a estimativa da média amostral estratificada

4) Obter a estimativa do desvio padrão da média amostral estratificada

5) Obter um Intervalo com 95 % de Confiança para  $\mu$

## Exercício

- 1) Obter a estimativa da proporção de pessoas favoráveis a campanha
- 2) Obter a estimativa da variância da proporção de pessoas favoráveis a campanha
- 3) Obter um Intervalo com 95 % de Confiança para  $\pi$