Bioestatística Básica

Curso de Pós-Graduação
RCA 5804

PROF. DR. ALFREDO J RODRIGUES

DEPARTAMENTO DE CIRURGIA E ANATOMIA

FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

UNIVERSIDADE DE SÃO PAULO

alfredo@fmrp.usp.br

Bioestatística Básica
RCA 5804

Proposta do curso

Oferecer conceitos básicos sobre os testes mais frequentemente utilizados em pesquisa biomédica

- Conceitos essenciais em estatísticas
- Quais são os testes mais comumente utilizados
- Quando utiliza-los
- Condições mínimas para aplica-los

Journal of Medical Systems, Vol. 27, No. 3, June 2003 (© 2003)

Methodological and Statistical Techniques: What Do Residents Really Need to Know About Statistics?

James F. Reed III,[1,3] Philip Salen,[2] and Pooneh Bagher[1]

Table IV. Statistical Methods by Medical Specialty

	Family Practice	Emergency Medicine	Obstetrics and Gynecology	*p*-value
t-tests	108 (28.4%)	196 (29.4%)	304 (38.4%)	0.001
χ^2	181 (47.6%)	312 (46.8%)	379 (47.9%)	0.918
OR	61 (16.1%)	81 (12.2%)	151 (19.1%)	0.002
ROC	3 (0.8%)	20 (3.0%)	29 (3.7%)	0.020
Kappa	10 (2.6%)	57 (8.6%)	10 (1.3%)	0.001
Survival	3 (0.8%)	15 (2.3%)	25 (3.2%)	0.042
Sensitivity	17 (4.5%)	61 (9.2%)	50 (6.3%)	0.011
Regression	80 (21.1%)	98 (14.7%)	145 (18.3%)	0.027
Nonparametric	27 (7.1%)	136 (20.4%)	223 (28.2%)	0.001
ANOVA	66 (17.4%)	122 (18.3%)	195 (24.7%)	0.002

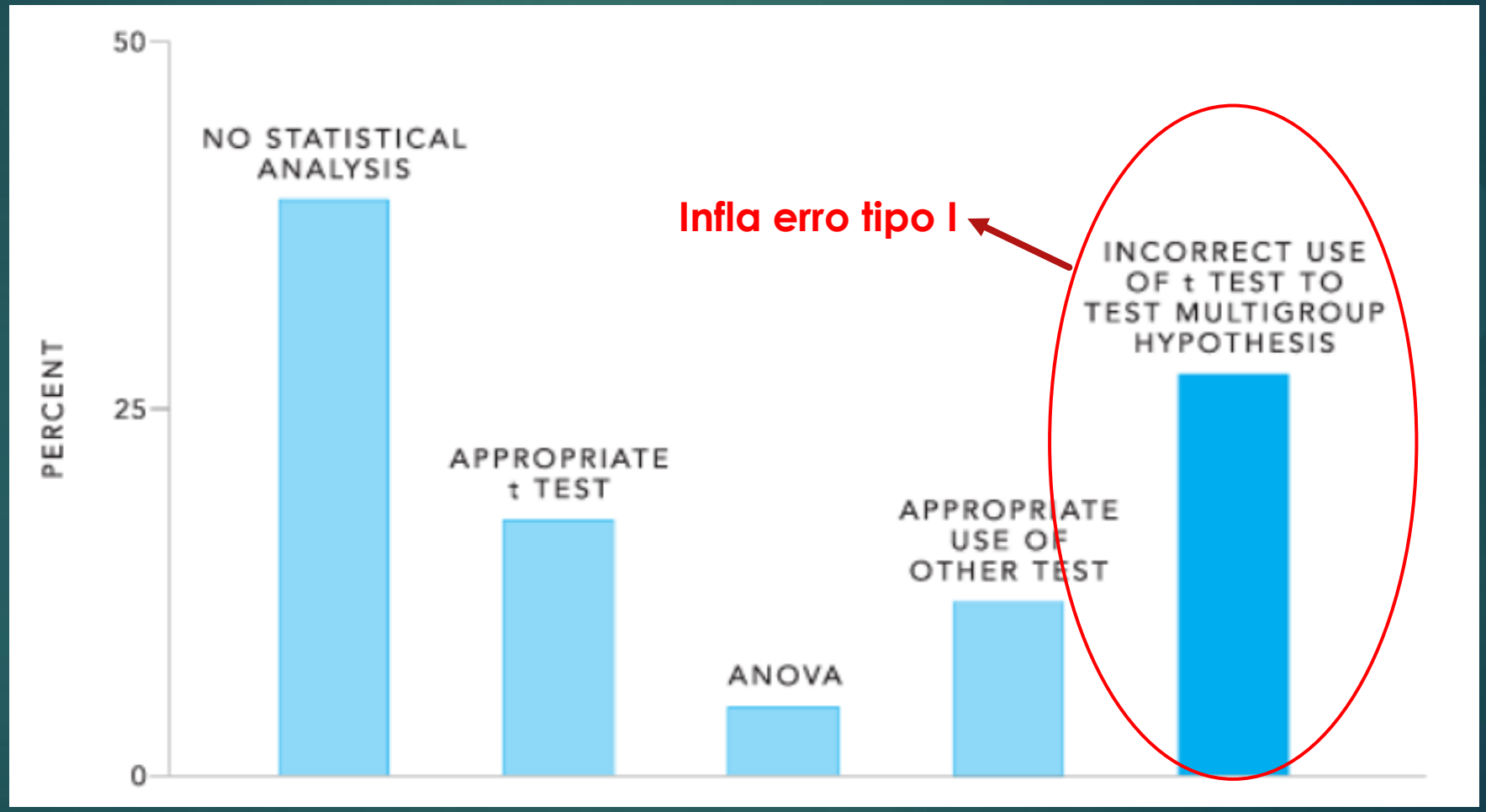
These results show that a physician who comfortably comprehends the appropriate use of descriptive statistics, Student's t-test, Pearson's chi-square/Fisher's Exact test will be able to read and interpret at least 70% of the published medical literature. Educational efforts should focus on appropriate study design and analysis.

Statistical errors in medical research – a review of common pitfalls

Incompatibility of statistical test with type of data examined
Unpaired tests for paired data or vice versa
Inappropriate use of parametric methods
Use of an inappropriate test for the hypothesis under investigation
Inflation of Type I error
Failure to include a multiple-comparison correction
Inappropriate post-hoc Subgroup analysis
Typical errors with Student's t-test
Failure to prove test assumptions
Unequal sample sizes for paired t-test
Improper multiple pair-wise comparisons of more than two groups
Use of an unpaired t-test for paired data or vice versa
Typical errors with χ^2-tests
No Yates-continuity correction reported if small numbers
Use of chi-square when expected numbers in a cell are <5
No explicit statement of the tested Null-Hypotheses
Failure to use multivariate techniques to adjust for confounding factors

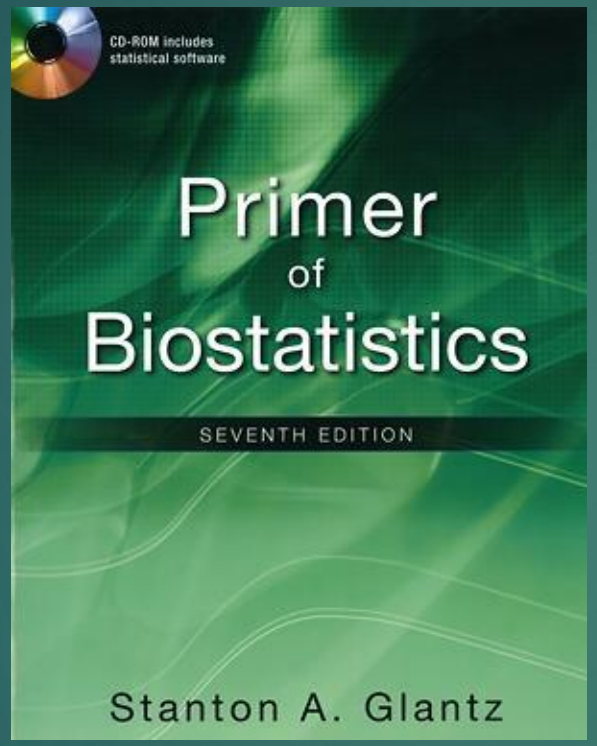
SWISS MED WKLY 2007;137:44–49 · www.smw.ch

USO INCORRETO DE TESTES ESTATÍSTICO

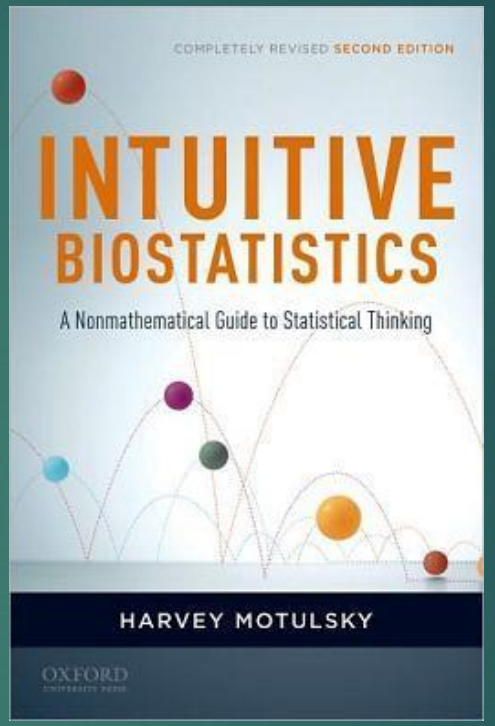


Glantz. Circulation. 61,1980

Sugestões de Bibliografia

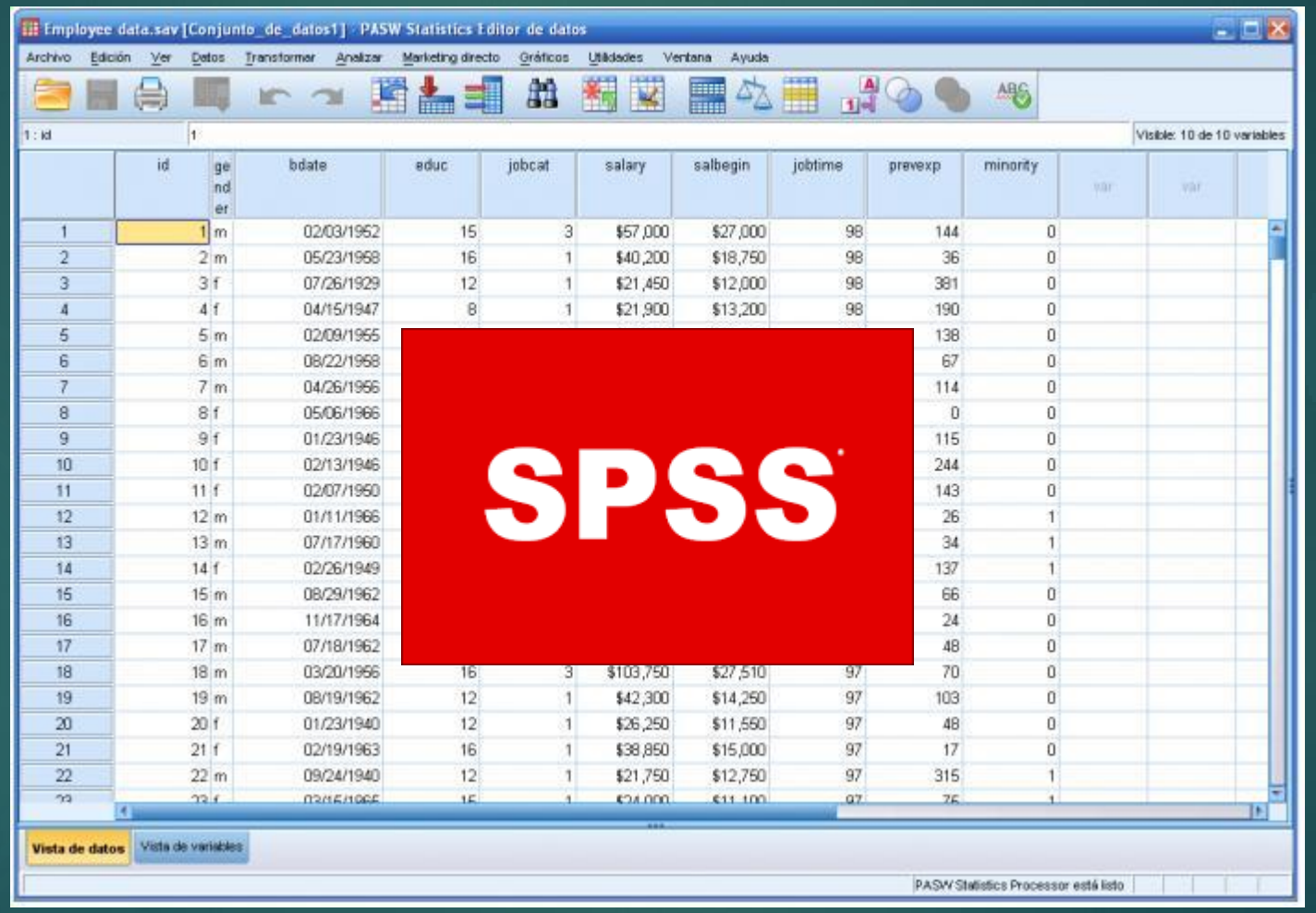


www.amazon.com.br

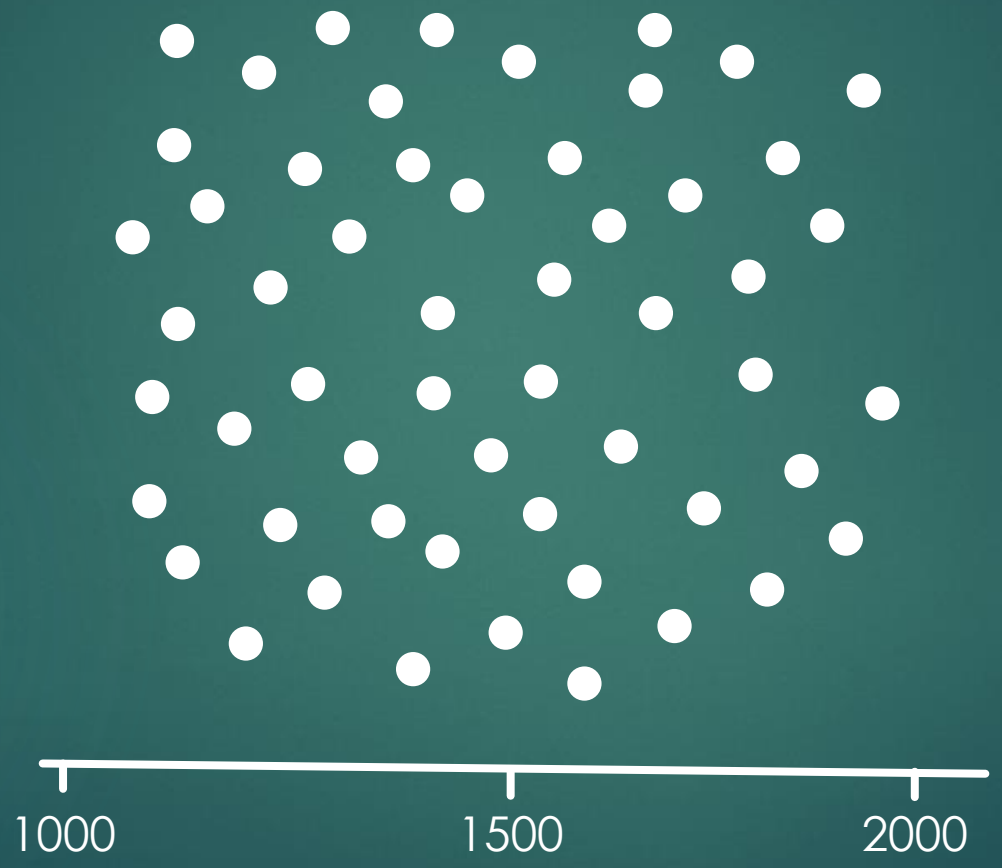


www.amazon.com.br

Software

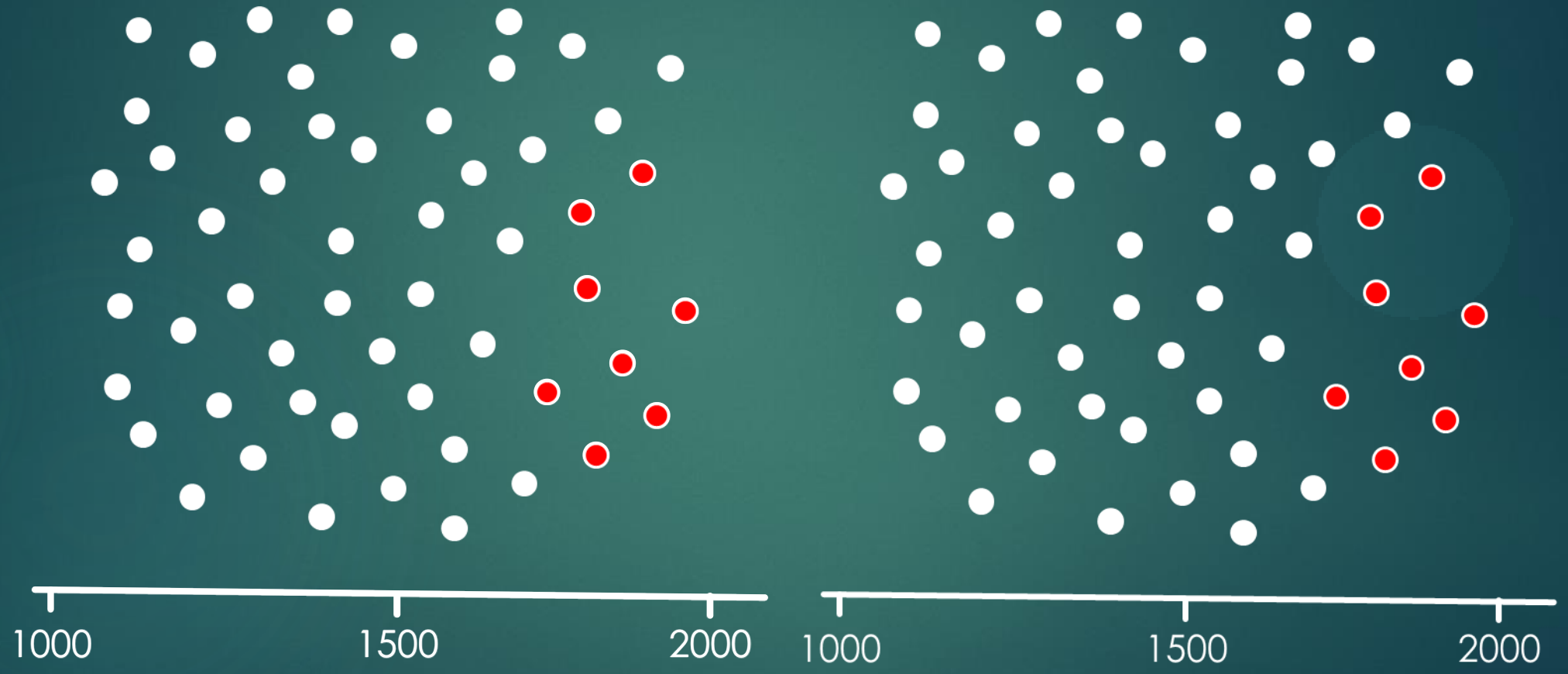

Employee data.sav [Conjunto_de_datos1] - PASW Statistics Editor de datos
Archivo Edición Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades Ventana Ayuda
1 : id
Visible: 10 de 10 variables
id gender bdate educ jobcat salary salbegin jobtime prevexp minority
1 1 m 02/03/1952 15 3 $57,000 $27,000 98 144 0
2 2 m 05/23/1958 16 1 $40,200 $18,750 98 36 0
3 3 f 07/26/1929 12 1 $21,450 $12,000 98 381 0
4 4 f 04/15/1947 8 1 $21,900 $13,200 98 190 0
5 5 m 02/09/1955 138 0
6 6 m 08/22/1958 67 0
7 7 m 04/26/1956 114 0
8 8 f 05/06/1966 0 0
9 9 f 01/23/1946 115 0
10 10 f 02/13/1946 244 0
11 11 f 02/07/1950 143 0
12 12 m 01/11/1966 26 1
13 13 m 07/17/1960 34 1
14 14 f 02/26/1949 137 1
15 15 m 08/29/1962 66 0
16 16 m 11/17/1964 24 0
17 17 m 07/18/1962 48 0
18 18 m 03/20/1956 16 3 $103,750 $27,510 97 70 0
19 19 m 08/19/1962 12 1 $42,300 $14,250 97 103 0
20 20 f 01/23/1940 12 1 $26,250 $11,550 97 48 0
21 21 f 02/19/1963 16 1 $38,850 $15,000 97 17 0
22 22 m 09/24/1940 12 1 $21,750 $12,750 97 315 1
SPSS
Vista de datos
Vista de variables
PASW Statistics Processor está listo

Volume urinário de cada membro de uma população



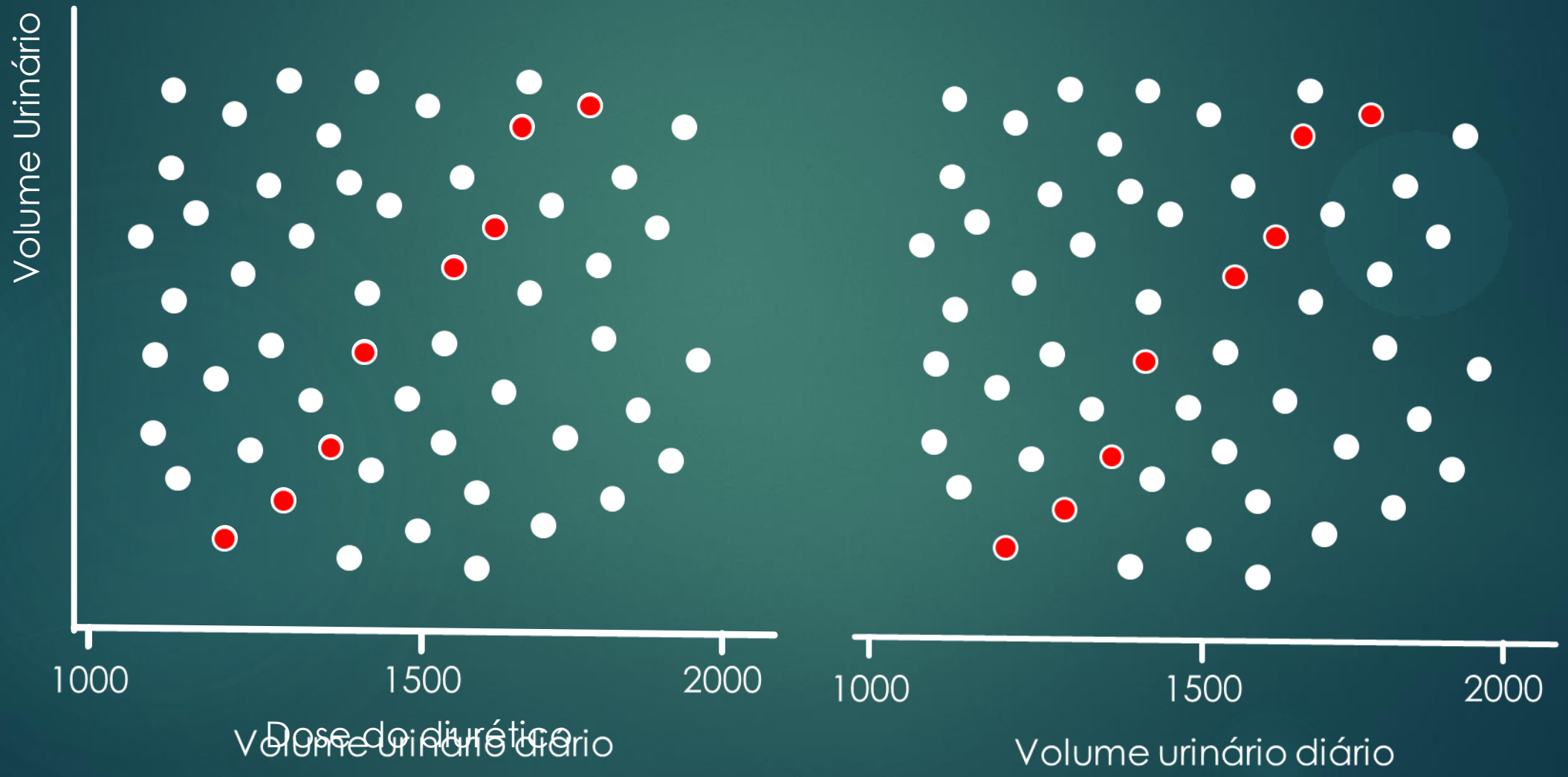
Volume urinário diário

Volume urinário uma amostra da população após uso de uma "droga diurética"



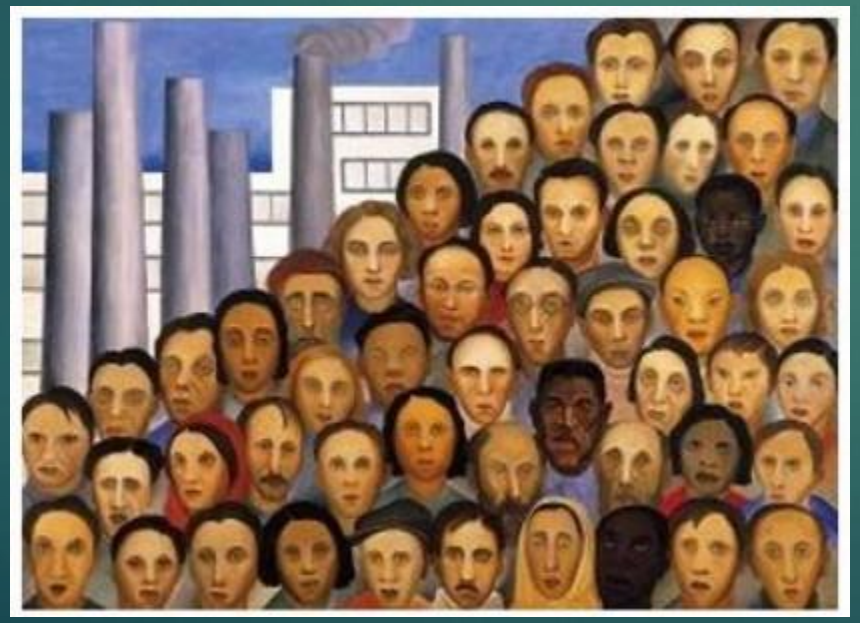
Volume urinário diário

Volume urinário com doses crescentes de uma "droga diurética"



Objetivos do Pesquisador

Descrição da "população"



TESTAR HIPÓTESE (INFERÊNCIA)

SUMARIZANDO OS DADOS

DESCREVER A DISTRIBUIÇÃO DE UM DADO OU VARIÁVEL DE INTERESSE EM UMA POPULAÇÃO DE FORMA A FORNECER UM PANORAMA DA POPULAÇÃO

Tipos de Variáveis

1. Variáveis Quantitativas

- **Discretas: Tanto ordem como magnitude importam**
  - Quantidade: numerous inteiros positivos
  - A ≠ entre dois valores é constante
    - Ex: número de filhos

- **Contínuas:**
- **Valores fracionais são possíveis**
- **1,24; 1,27; 2,0; 2,3; 3,15**
  - Ex: nível colesteral, peso, glicemia

Tipos de Variáveis

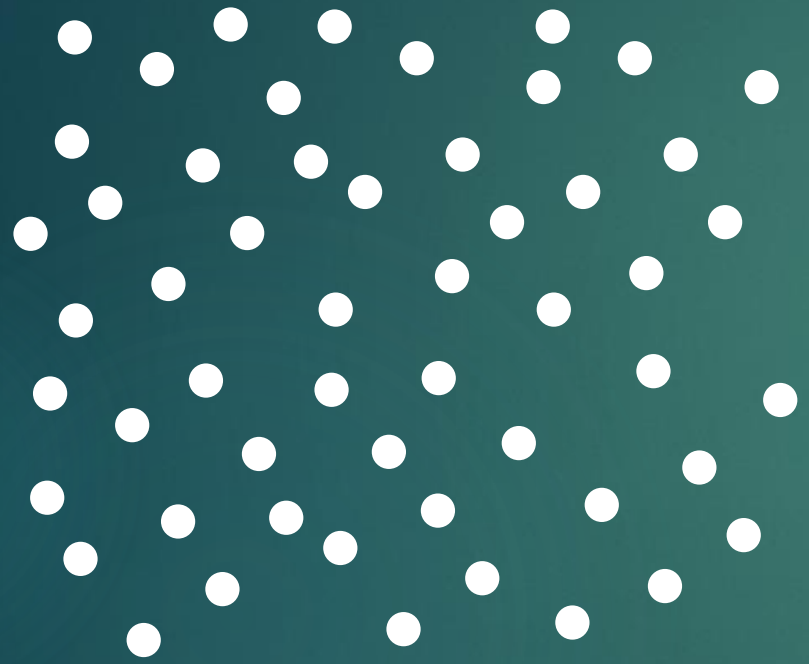
2. Variáveis Categóricas

- **Nominal:** ordem e magnitude não importa
  - sexo, raça,
  - Duas (binários ou dicotômicos) ou + categorias
- **Ordinal:** há "magnitude" e ordem importa
  - classe funcional NYHA, nível escolaridade

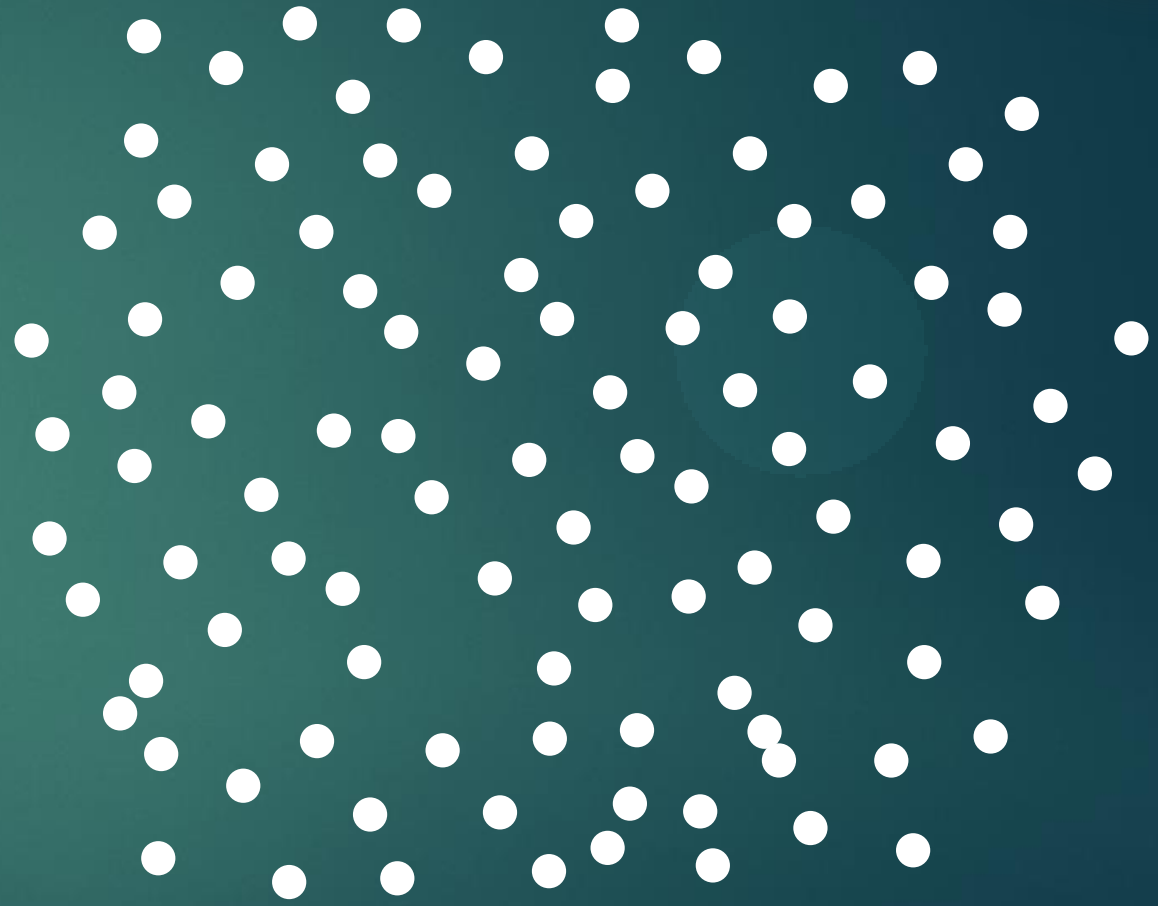
•A ≠ de magnetude entre as categorias não é obrigatoriamente a mesma

•Ex: leve ⟷ moderada ↔ grave

População 1



População 2



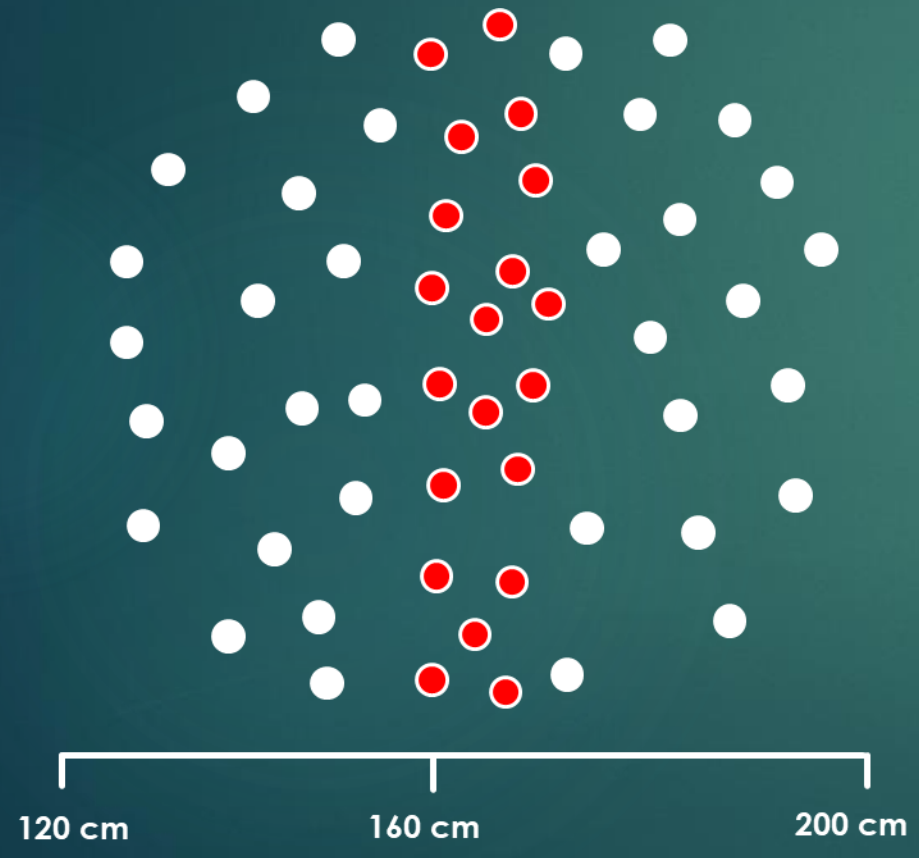
Tamanho População 1 < tamanho População 2

Estatura de cada membro de uma população

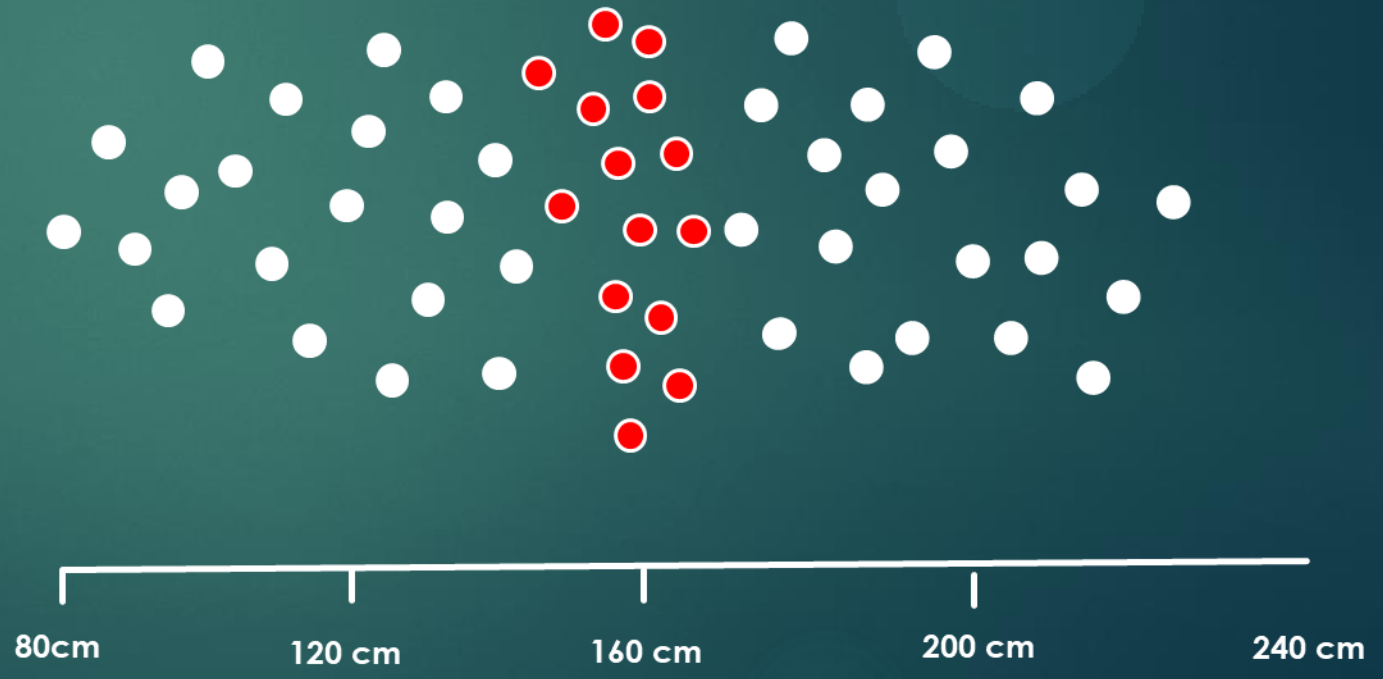
Tamanho da população 1 = tamanho da população 2

Ambas as populações se concentram em torno de 160 cm

Variação de estatura na população 1 < que na população 2



População 1



População 2

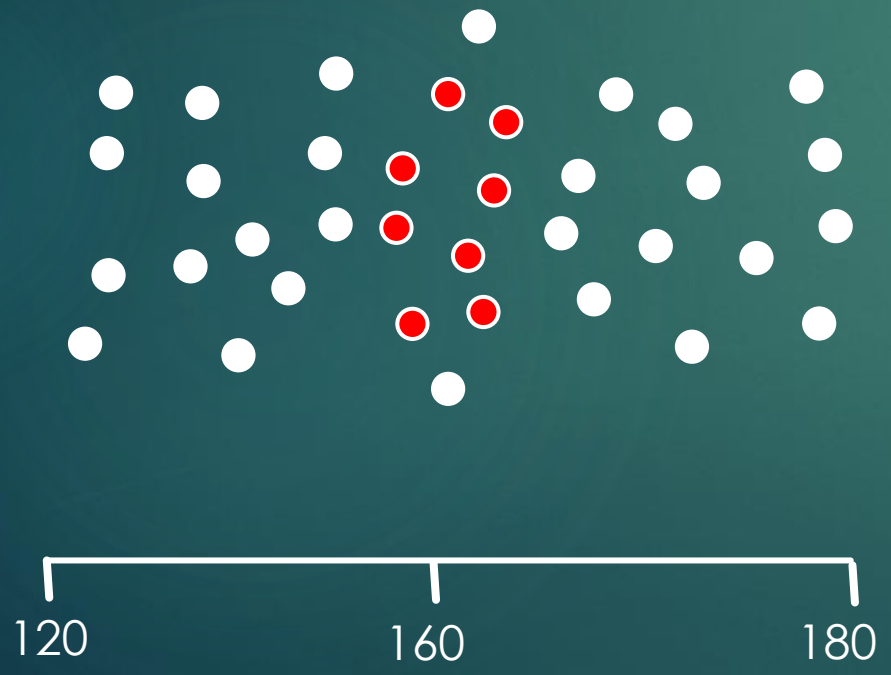
Estatura de cada membro de uma população

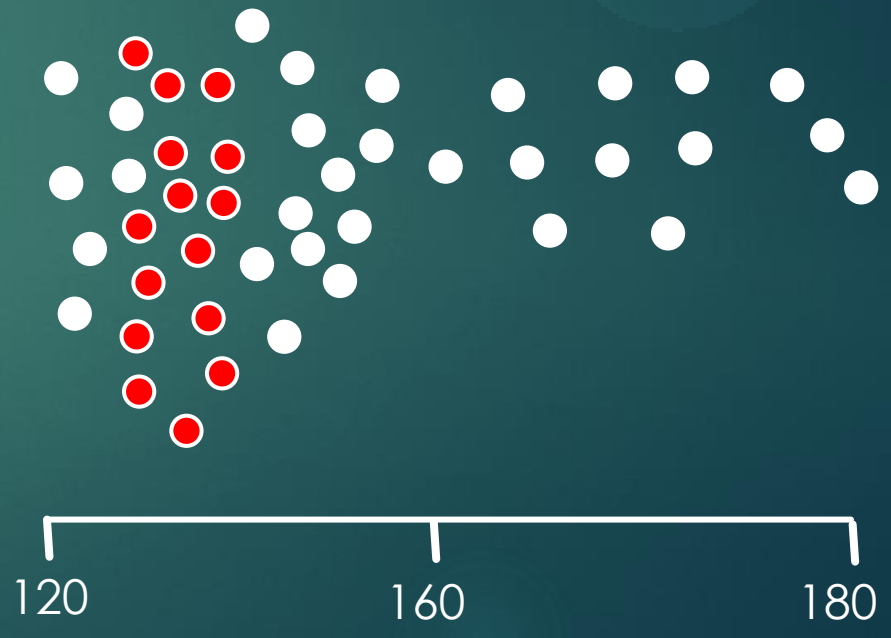
Tamanho da população 1 = tamanho da população 2

Ambas as populações se concentram em torno de 160 cm

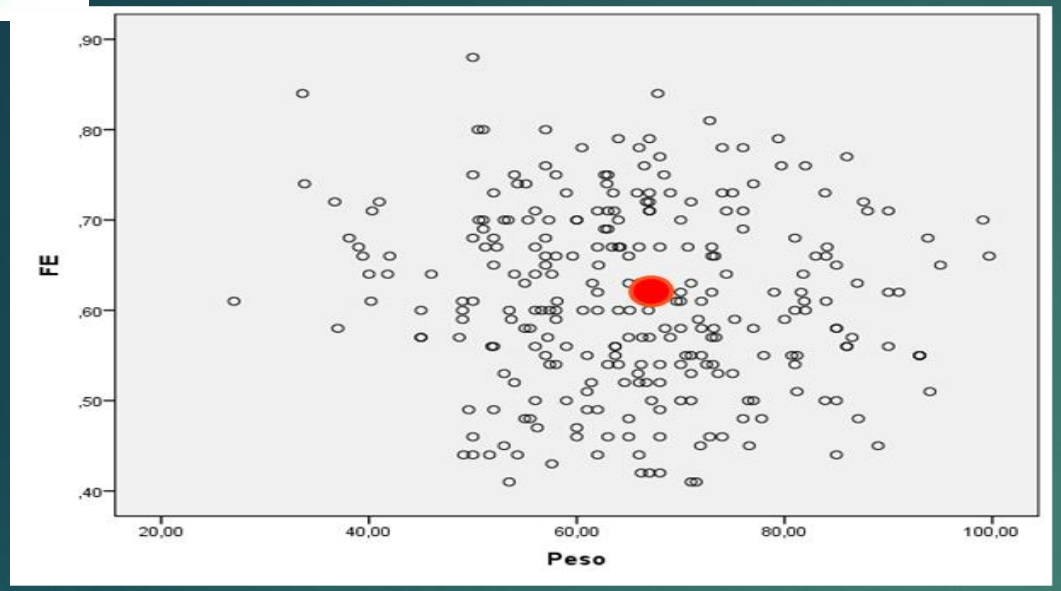
Variação de estatura na população 1 = que na população 2

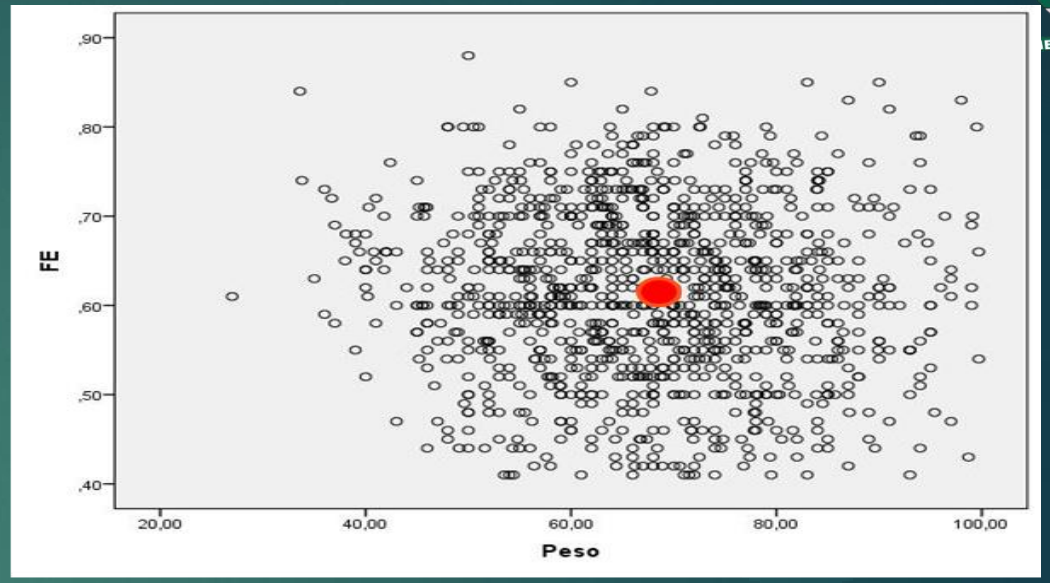
Forma da distribuição de estatura na população 1 ≠ da população 2

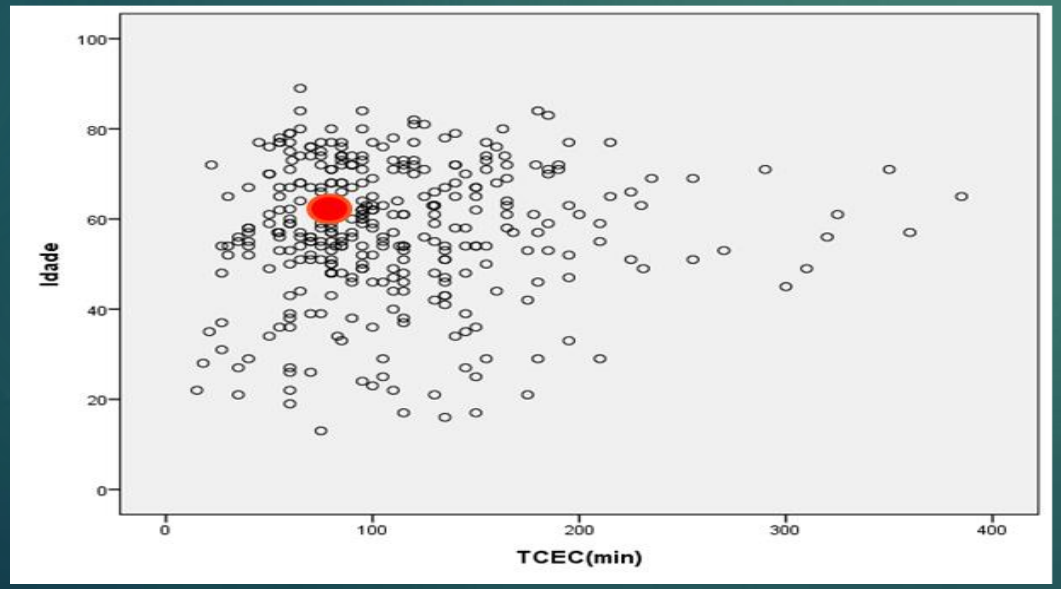


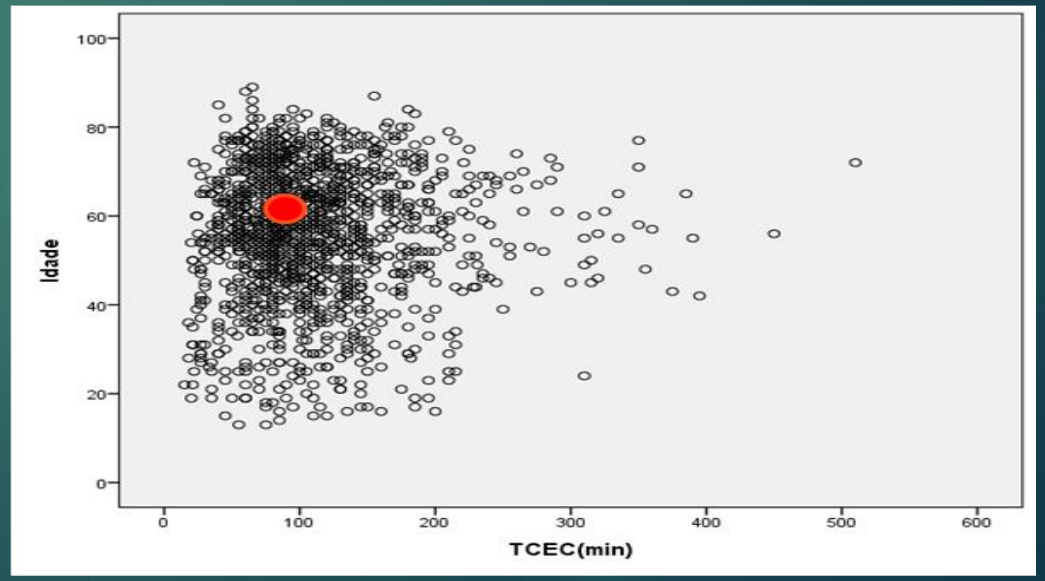


Tamanho (n), média e distribuição forma da distribuição









ESTATÍSTICA DESCRITIVA

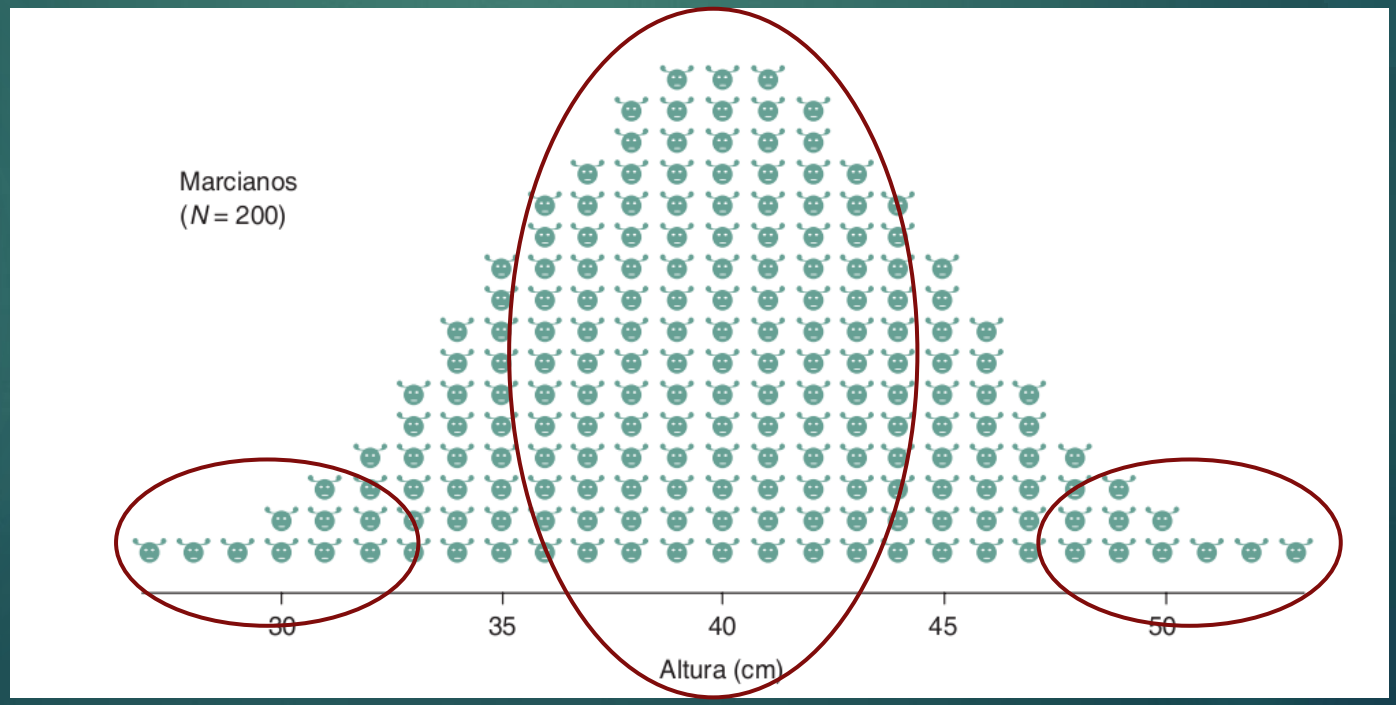
Os 4 principais descritores (parâmetros) populacionais:

1. O **tamanho** da população (n)
2. Uma medida de "**tendência central**" (média)
3. Uma medida de **dispersão ou variação** em torno deste valor central (variância)
4. **A forma** como a variável de interesse esta distribuída ao redor do centro

*** As 3 primeiras só descrevem corretamente a população se a distribuição for simétrica**

Forma da Distribuição
Distribuição simétrica (Normal ou "gaussiana")

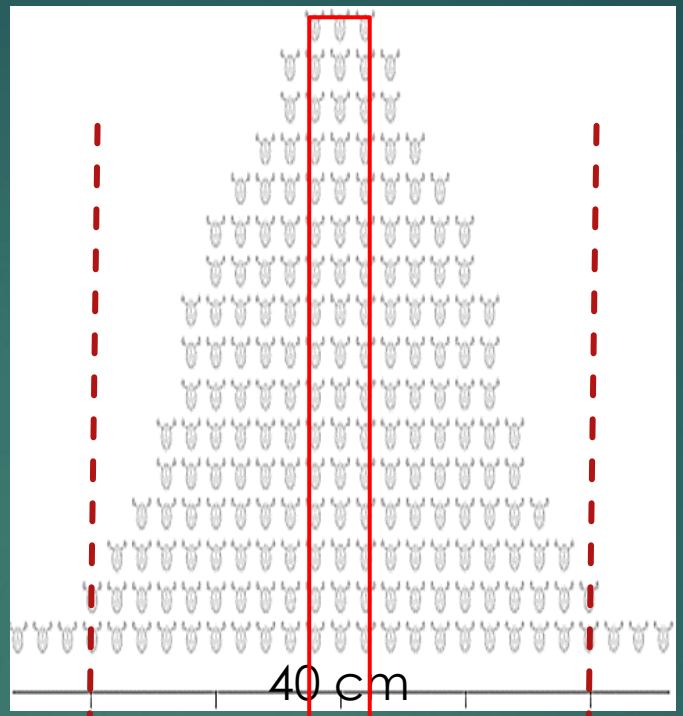
- Probabilidade de q.q. indivíduo estar próximo ao centro é MAIOR que a probabilidade de estar nas extremidades
- Probabilidade de q.q indivíduo estar a direita do centro **é IGUAL** a probabilidade de estar a esquerda e vice-versa



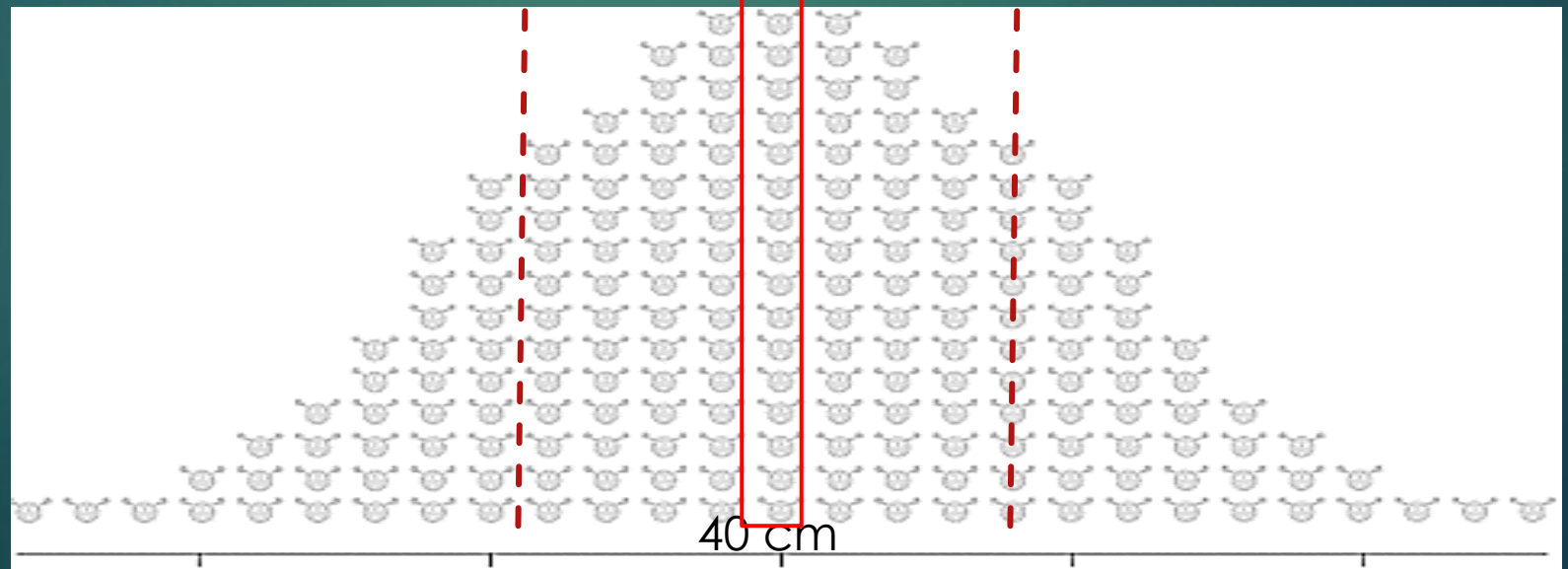
Forma simétrica da dispersão a partir do do centro

μ = media

N1=N2=200
μ1= μ2=40cm



N1=N2=200
μ1= μ2=40cm



Forma assimétrica da dispersão a partir do do centro

Probabilidade de q.q indivíduo estar em uma das extremidades é ≠ da probabilidade de estar na outra extremidade

Medidas de tendência Central

Medidas de Tendência Central

- **Média**
- **Mediana**
- Moda

Média aritmética ou média

$$\bar{x} = \frac{\sum x_i}{n}$$

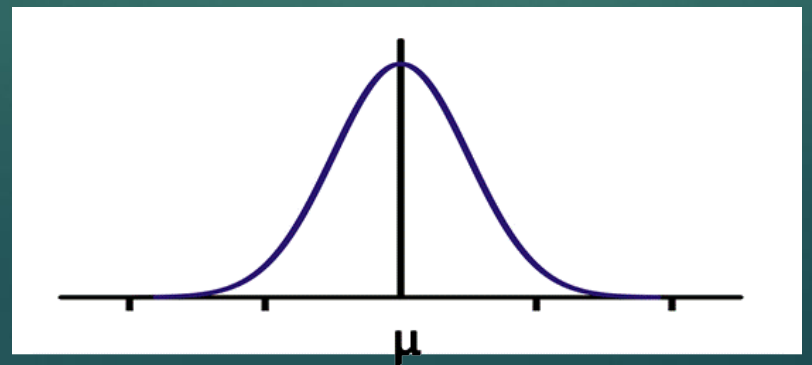
xi: valores individuais da amostra

X: valores individuais da população

n: número de valores da amostra

N: número de valores de uma população

A mais importante medida de tendência central, quando a distribuição é simétrica (normal ou gaussiana)



Média

Média amostral

$$\bar{x} = \frac{\sum x_i}{n}$$

Média populacional

$$\mu = \frac{\sum x}{N}$$

Mediana

Valor do meio do conjunto de dados, quando os valores estão dispostos em ordem (crescente ou decrescente)

- **Divide um conjunto de dados em duas partes iguais.**
- **Para calcular**
  - Disponha os valores em ordem (crescente ou decrescente)
  - Encontre a posição da mediana**: (n+1)/2**
    - **n= números de valores**
  - Se n é ímpar: **mediana é o valor correspondente á posição exatamente no meio**
  - Se n é par: mediana é a **MÉDIA entre os dois valores em torna da posição do meio**.

N é impar

Encontre a posição da mediana:
(n+1)/2

No exemplo: n=13 (impar)
Posição: (n+1)/2 = 7°
posição
Mediana: = 5

posição	valores
1	1,00
2	1,00
3	2,00
4	2,00
5	3,00
6	4,00
7	5,00
8	6,00
9	7,00
10	8,00
11	9,00
12	10,00
13	13,00

50%

50%

N é par

Encontre a posição da mediana: (n+1)/2

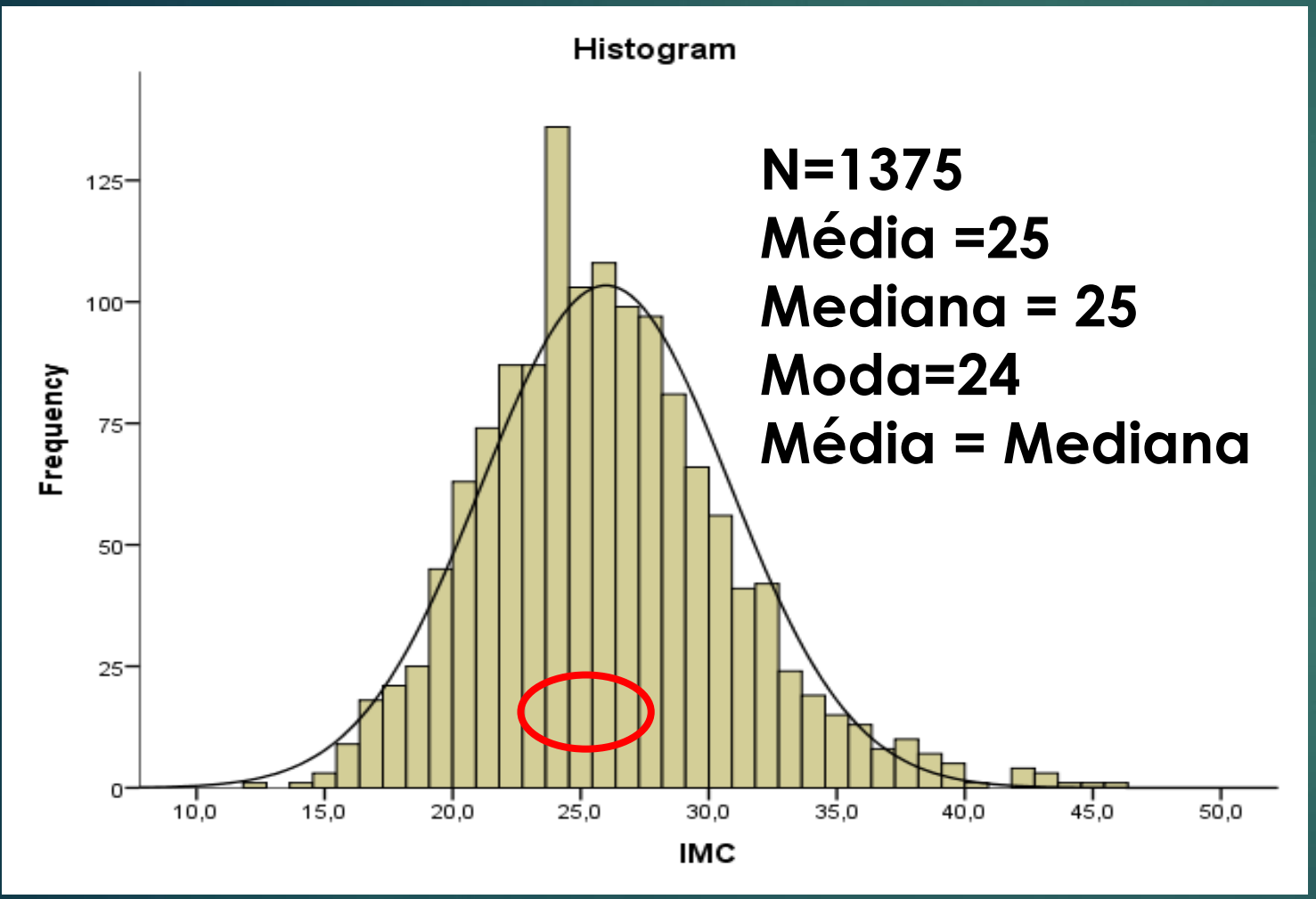
- No exemplo: n=12 (par)
  - Posição: (n+1)/2 = 6,5
  - Mediana= média entre o 6º e o 7º valores
  - Mediana= 4+5/2 = **4,5**

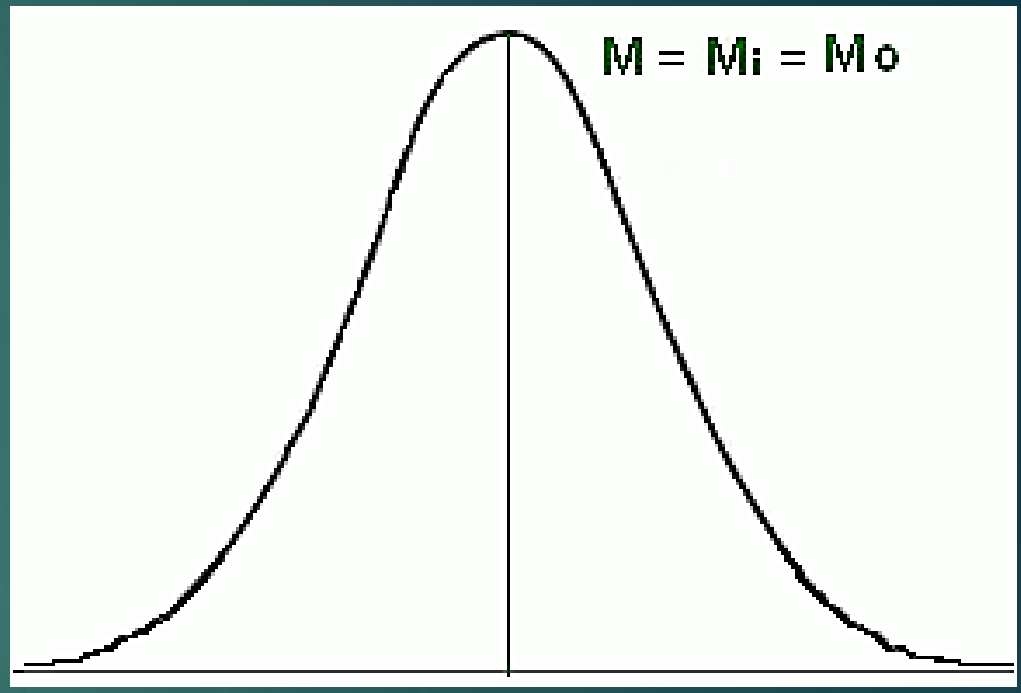
posição	valores
1	1,00
2	1,00
3	2,00
4	2,00
5	3,00
6	4,00
7	5,00
8	6,00
9	7,00
10	8,00
11	9,00
12	10,00

Moda

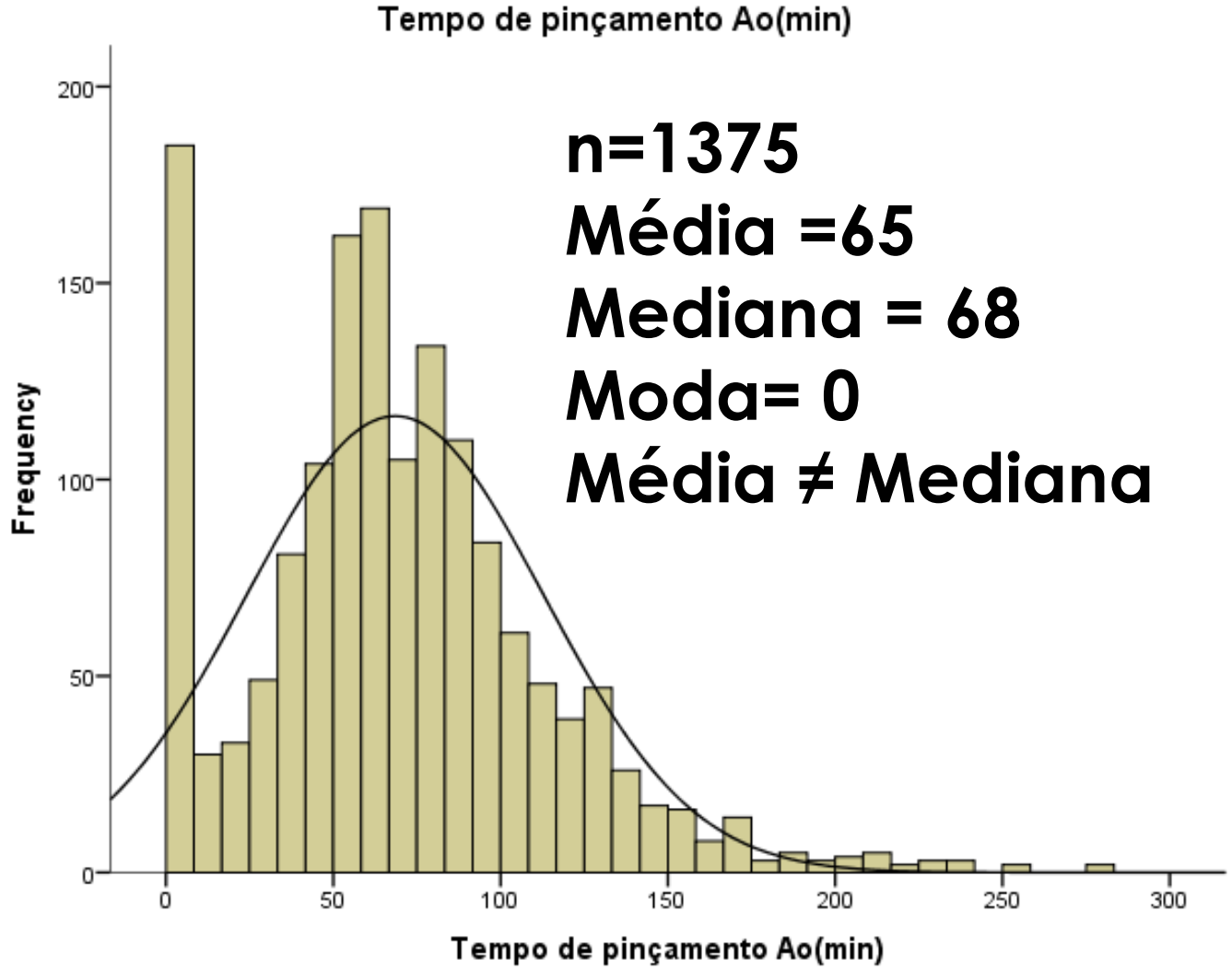
- É o valor que ocorre com maior frequência.
- Quando 2 valores ocorrem com a mesma frequência, cada um deles é chamado de uma moda, e o conjunto se diz **BIMODAL**
- Se mais de 2 valores ocorrem com a mesma frequência máxima, cada um deles é uma moda e o conjunto é **MULTIMODAL.**
- Quando nenhum valor é repetido o conjunto **não tem moda (AMODAL)**

Distribuição Normal

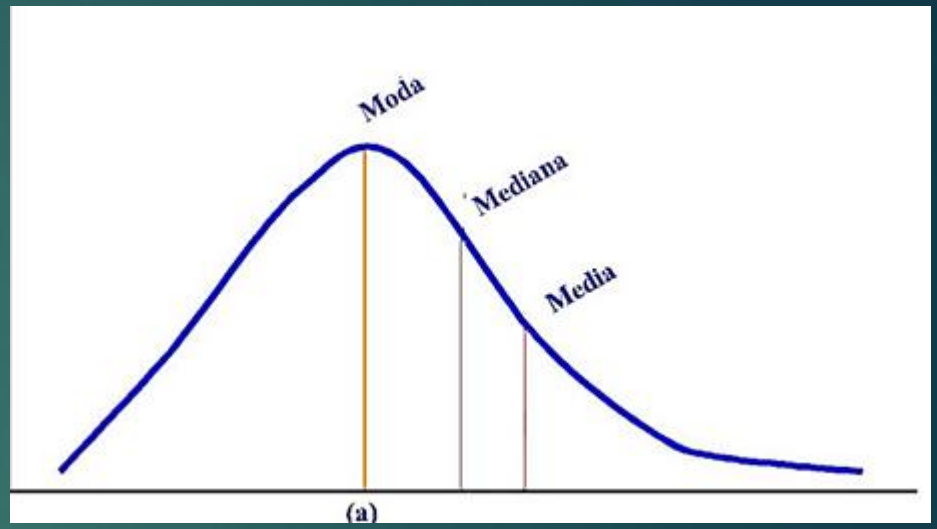


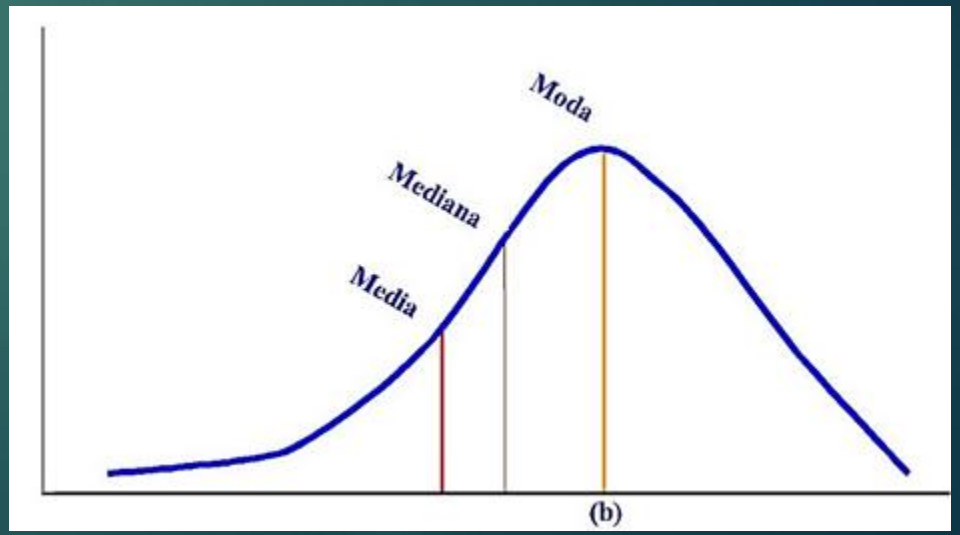


Distribuição não-Normal



n=1375
Média =65
Mediana = 68
Moda= 0
Média ≠ Mediana





Mediana x Média

- Seja o seguinte conjunto de 10 valores:

5, 7, 9, 11, 13, 15, 17, 19, 21, 23

n= 10

Média=14,0

Mediana=14,0

Moda-amodal

Mediana x Média

Alterando significativamente um dos valores

5, 7, 9, 11, 13, 15, 17, 19, 21, 110

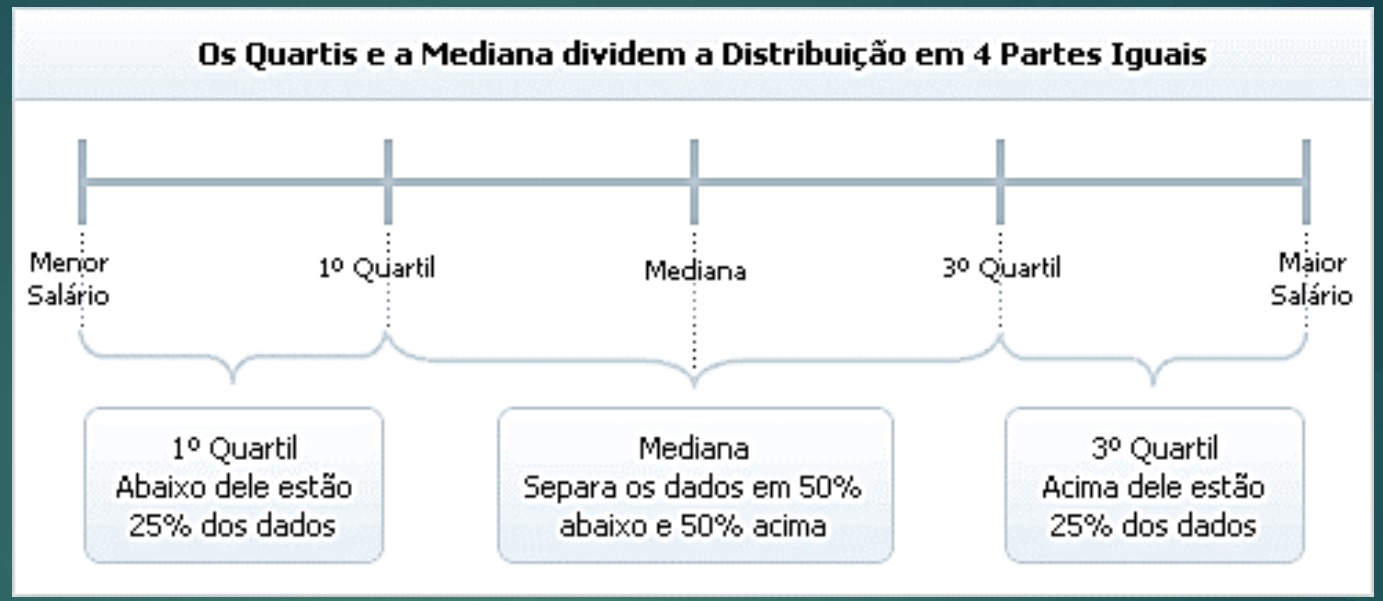
n= 10	n= 10
Média=14,0	**Média=22,7**
Mediana=14,0	Mediana=14,0
Moda-amodal	Moda- amodal

A MÉDIA É AFETADA POR VALORES EXTREMOS, A MEDIANA NÃO

Distribuição assimétrica- Quartil

$$Q_i = \frac{i}{4}(N+1)$$

i= quartil 1°, 2°, etc
N= número de valores



0,5; 0,7; 0,7; 0,9; 1,0; 1,1; 1,1; 1,2; 1,3; 1,3; 1,5; 1,8; 2,1; 2,2; 2,5; 2,5

n (número de valores)= 16 **valores ordenados**

Q1 (primeiro quartil): i=1; ¼ (16+1) = 0,25 x 17 = 4,24 (posição entre a 4° e 5°)
Toma-se a média entre o 4° e o 5° valores: Q1 = (0,9 + 1,0)/2 = 0,95.

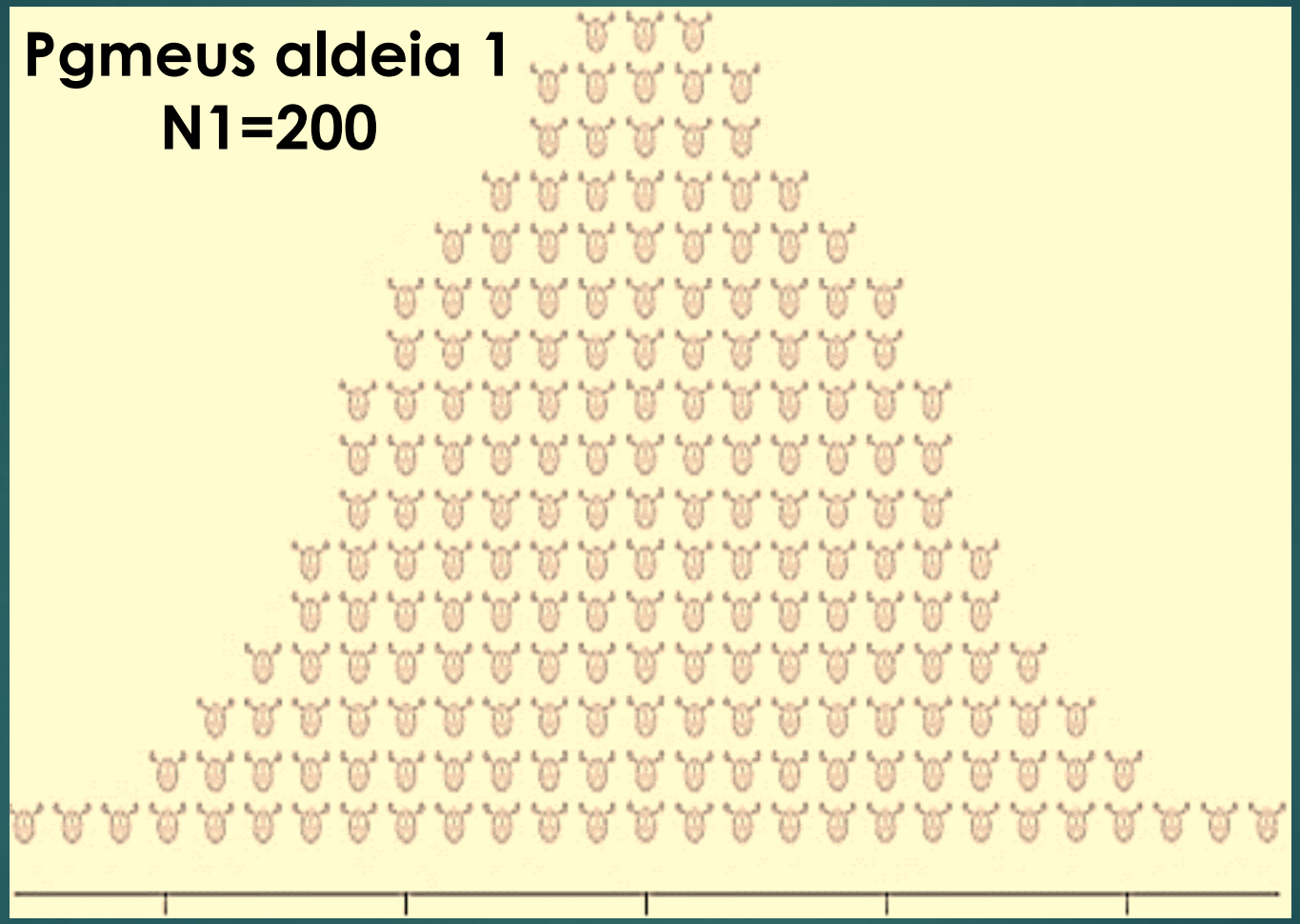
Q3 (terceiro quartil): i=3; 3/4 (16+1)= 12,75.
Toma-se a média entre o 12° e o 13° valores: Q3 = (1,8 + 2,1)/2 = 1,95.

Medidas de Tendência Central
Conclusões

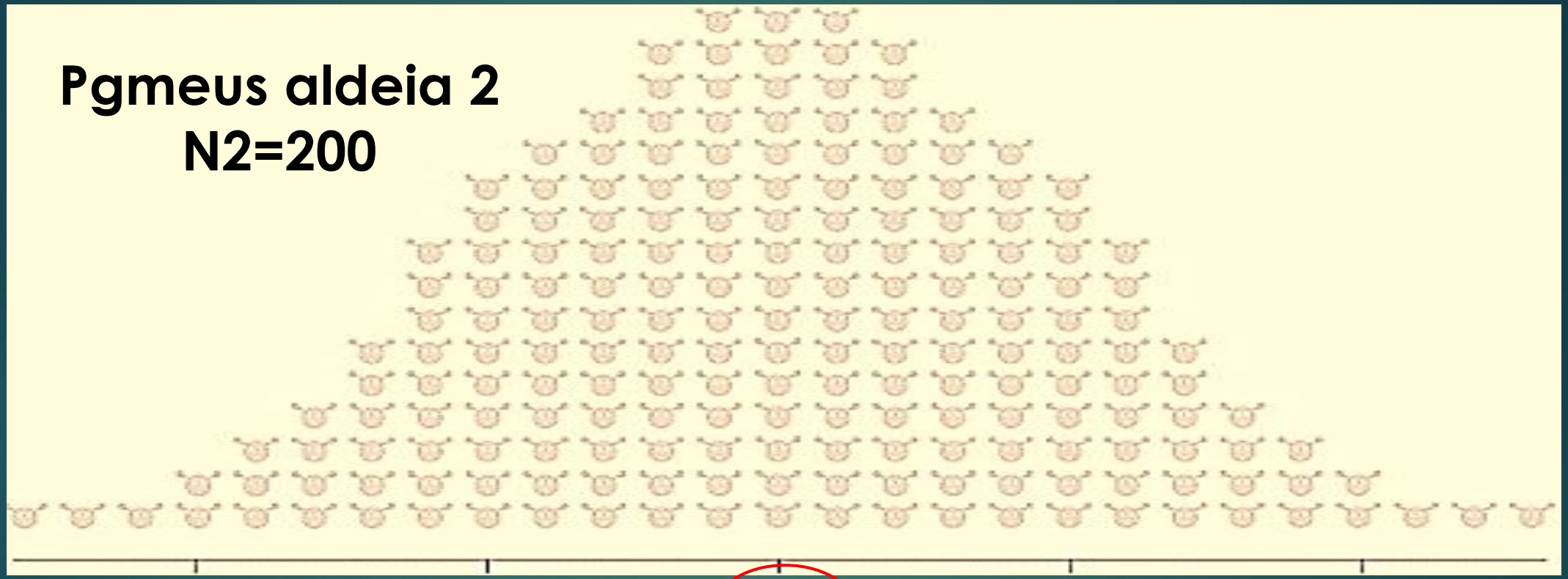
- Devemos ter cuidados ao escolhermos uma medida de posição para representar um conjunto de dados, pois:
  - A "*Média*" é afetada por valores extremos
  - Se a distribuição **não é simétrica a média não é uma medida de tendência central adequada**
  - Valores de "*Média*" e "*Mediana*" próximos é uma indicação que o conjunto de valores é razoavelmente simétrico em relação à posição central

Medidas de Dispersão

Distribuição Normal dos dados

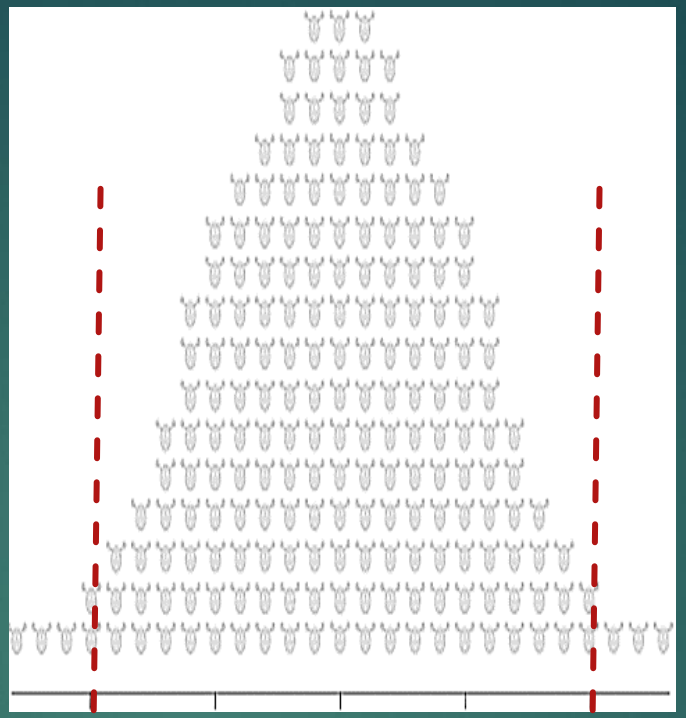


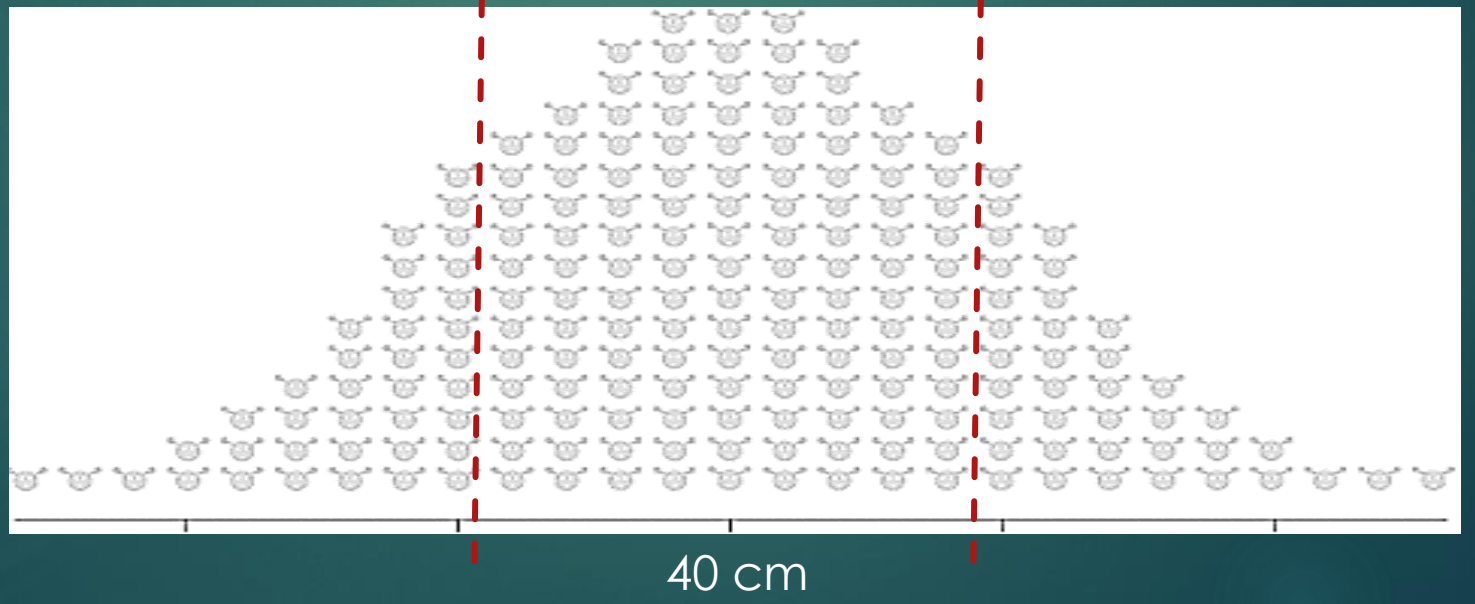
Distribuição Normal dos dados



Pgmeus

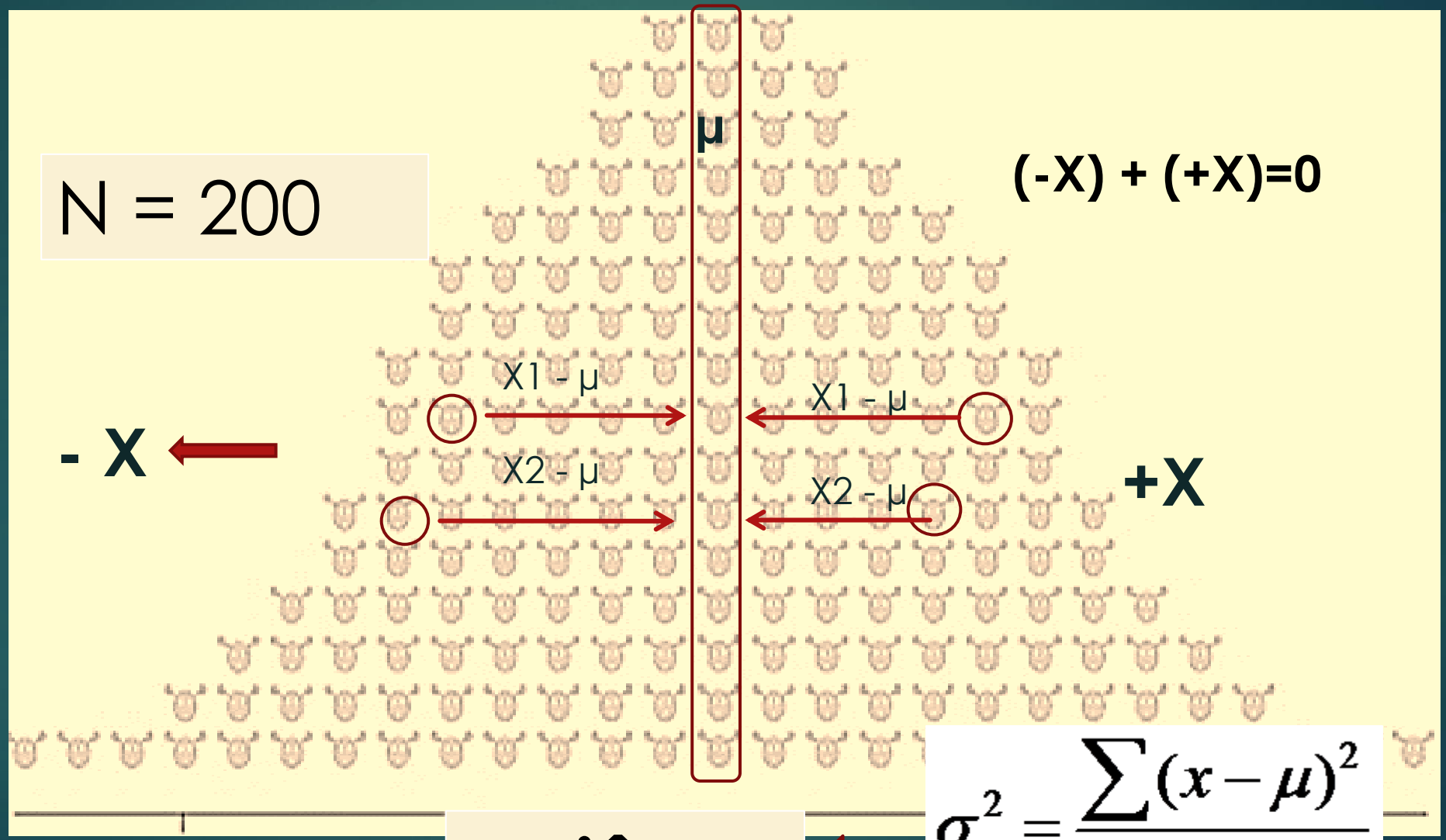
N1=N2=200
μ1= μ2=40cm





Medidas de Variabilidade

- Média da distância" de cada indivíduo da "media da população"



variância

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N}$$

Desvio-padrão

Variância

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Unidade ^{2}

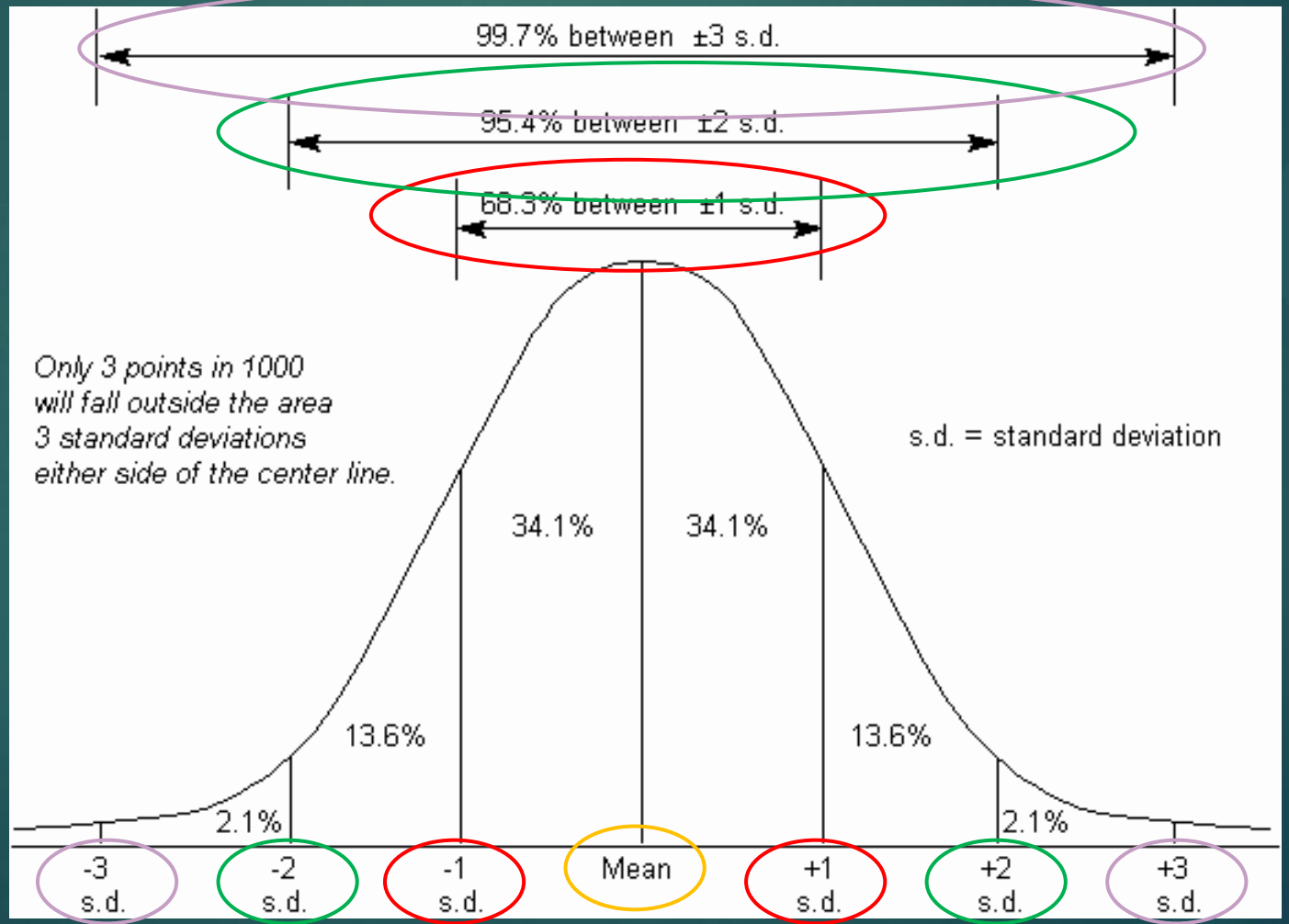
$\sqrt{}$

Desvio-padrão

$$\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{N}}$$

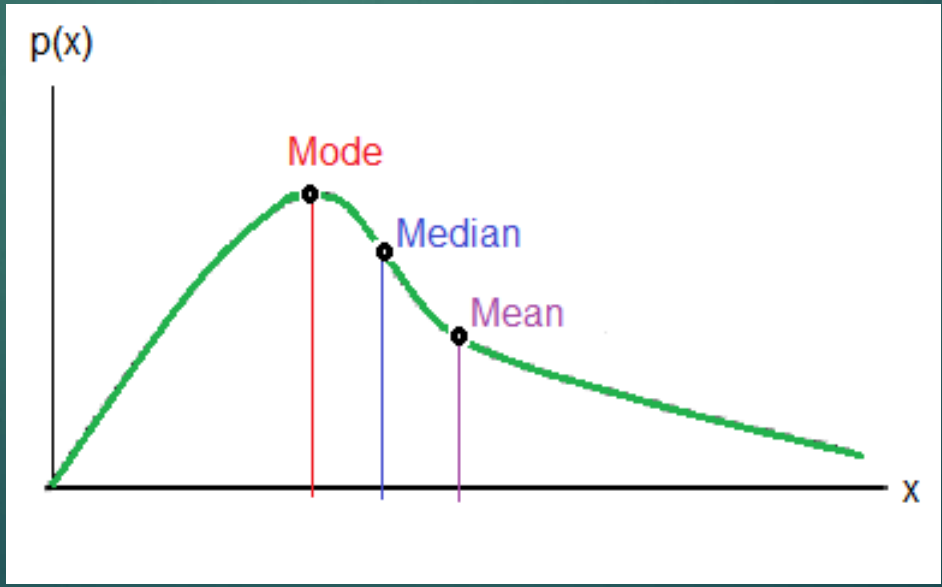
Unidade

Se distribuição é simétrica o DESVIO-PADRÃO estima a dispersão na amostra em torno da média



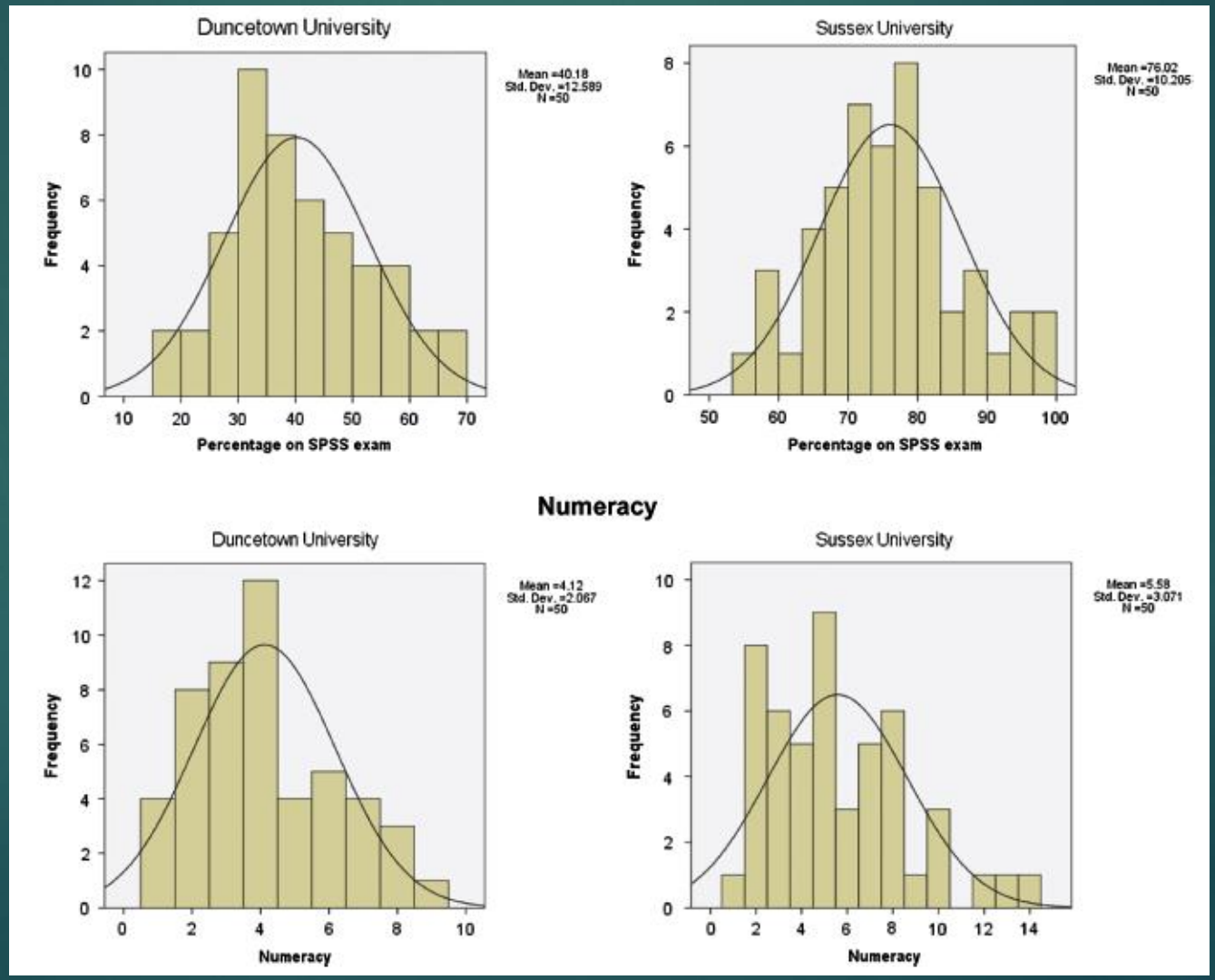
Distribuição Não-Normal (assimétrica)

- Média e desvio-padrão não são parâmetros descritivos dos dados com distribuição "assimétricas (não-normais)". Portanto, NÃO SÃO PARAMÉTRICOS (NÃO-PARAMÉTRICOS)



Verificação da Distribuição

- Histograma (distribuição de frequências)



Testes para verificação da distribuição

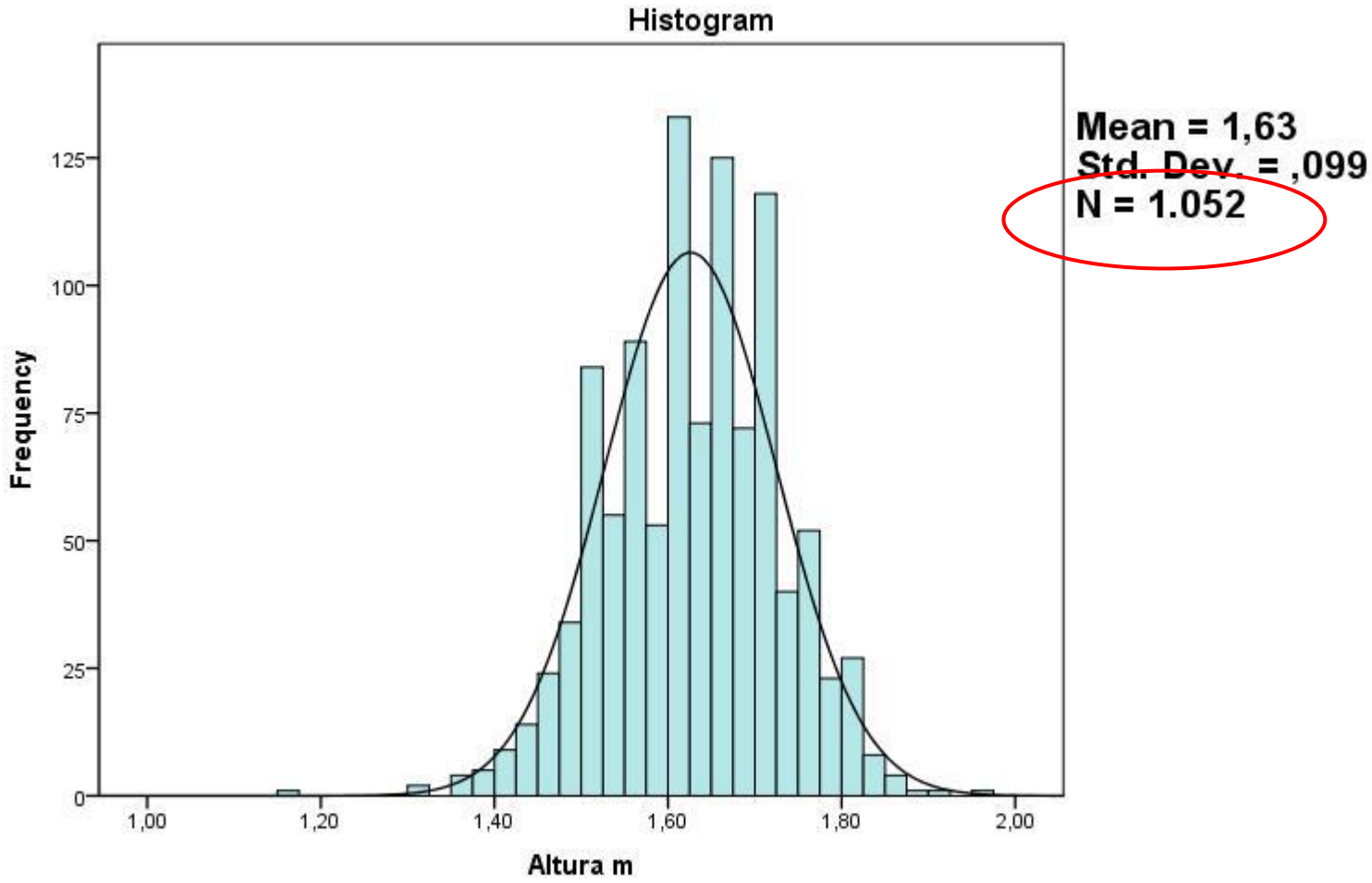
- D'Agostino-Pearson omnibus test
- Kolmogorov-Smirnov
- Shapiro-Wilk

} Qual a probabilidade da distribuição NÃO SER SIMÉTRICA

p>0,05 (≥5%) = distribuição SIMÉTRICA (NORMAL)
p<0,05 (<5%)= distribuição ASSIMÉTRICA (NÃO NORMAL)

Cuidado:

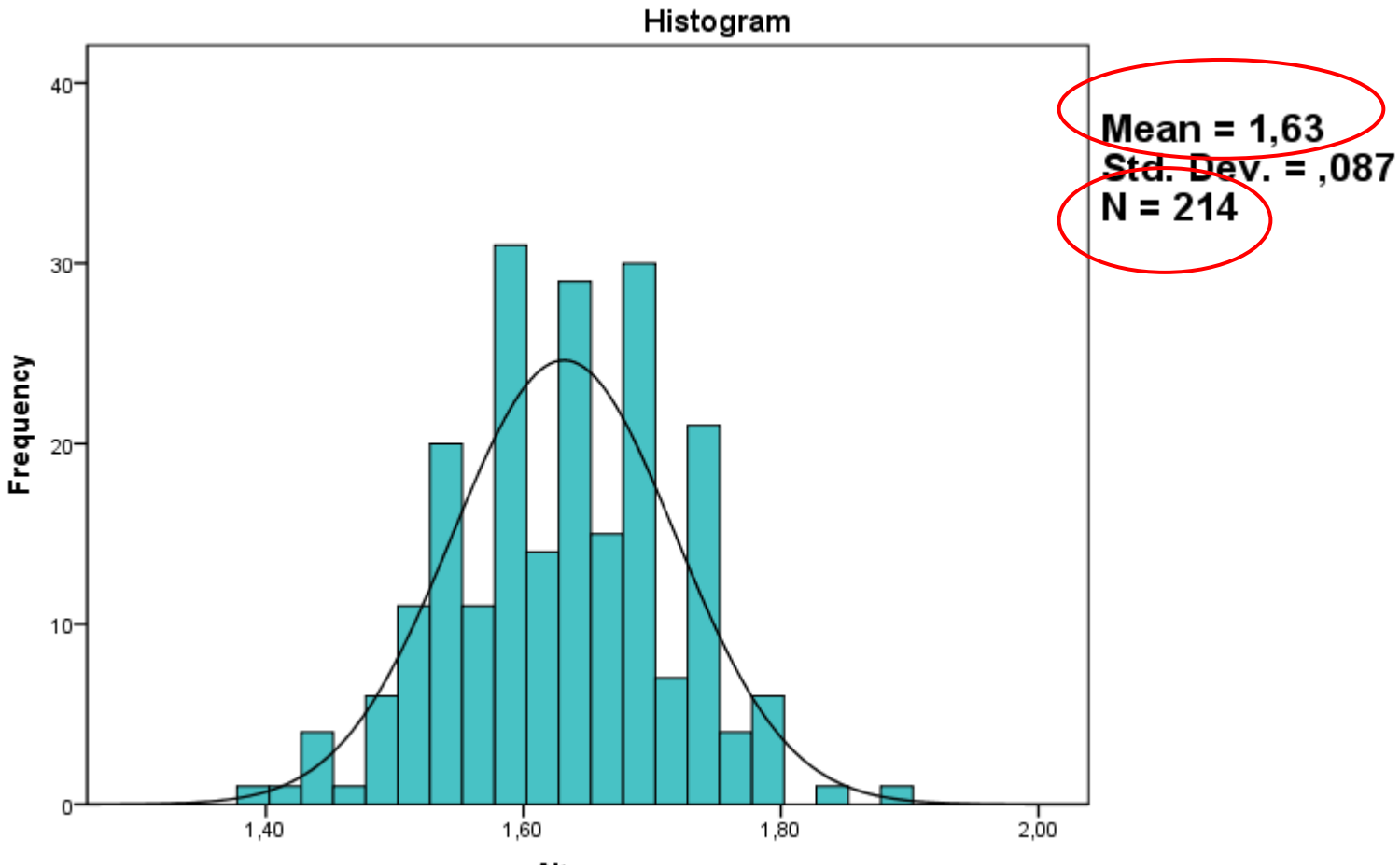
Em amostra grandes é possível se obter resultados significativos (p<0,05, assimétrica) em amostras com distribuição na verdade simétrica (normal) *(efeito do "n grande" no poder do teste em identificar desvio pequenos como significantes)*.



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Altura m	,045	1052	,000	,995	1052	,001

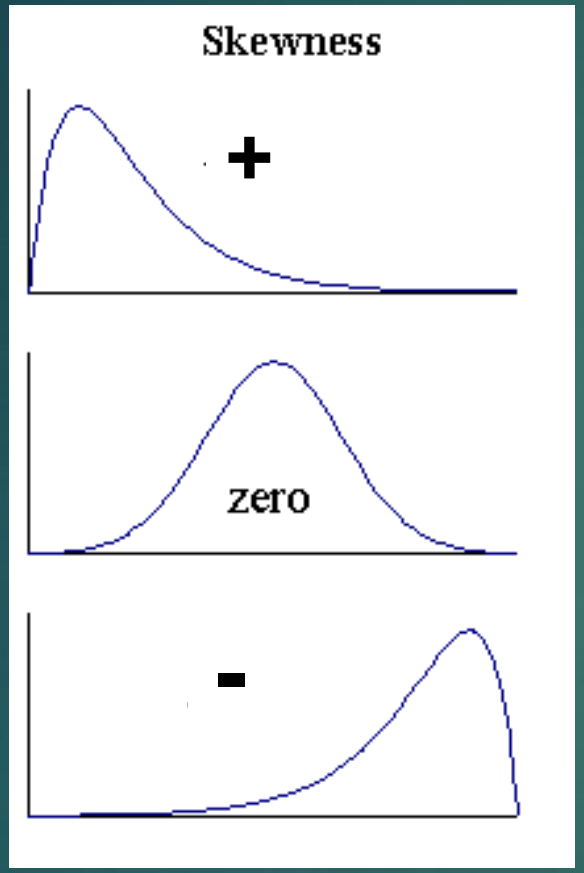
a. Lilliefors Significance Correction

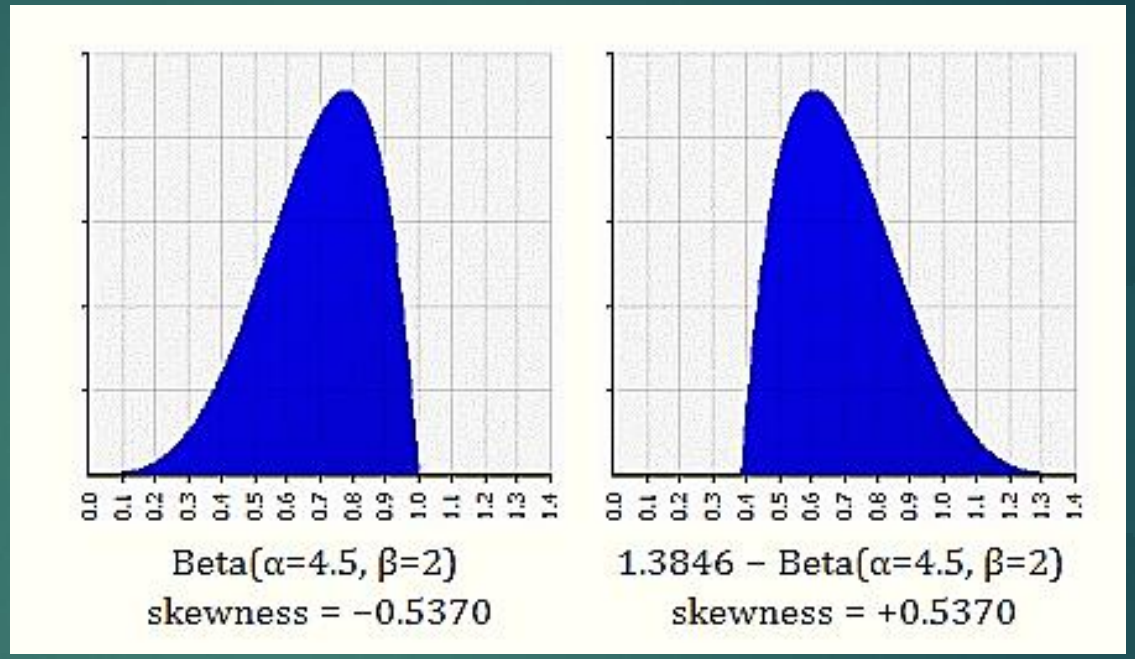


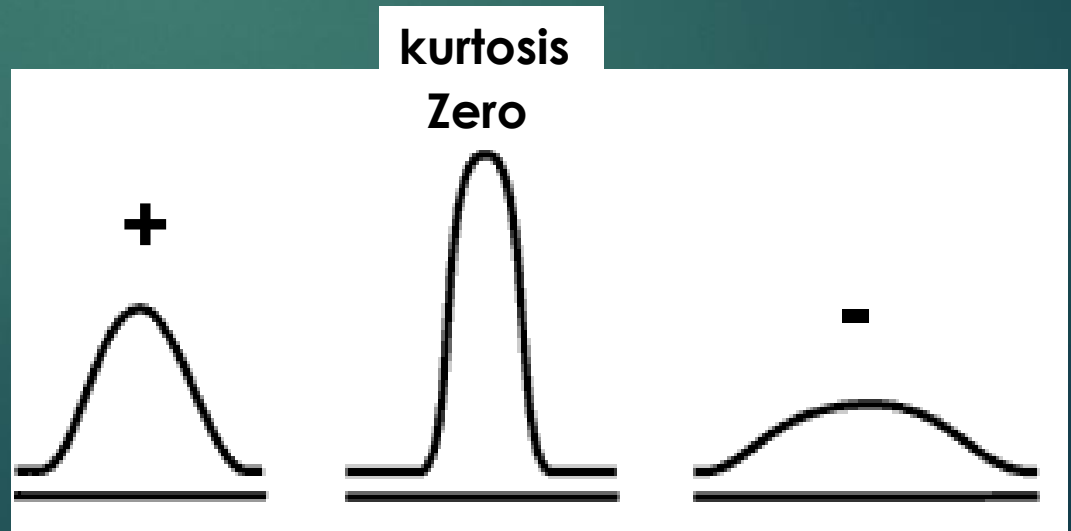
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Altura m	,047	214	,200*	,994	214	,590

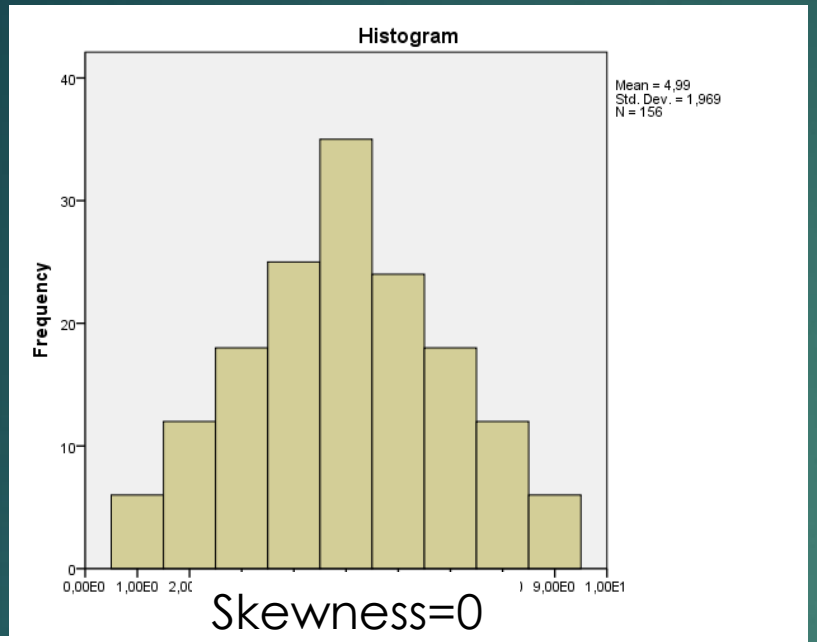
Verificação da Distribuição



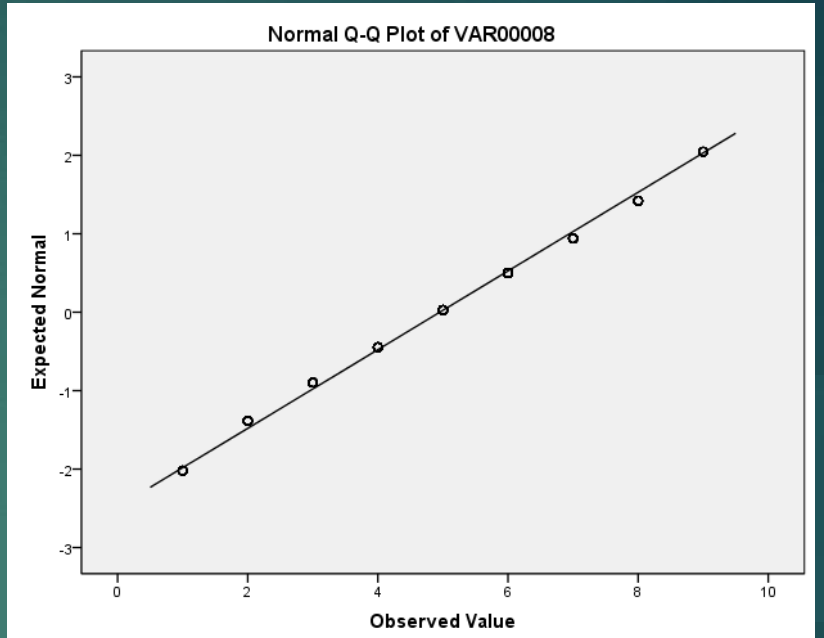


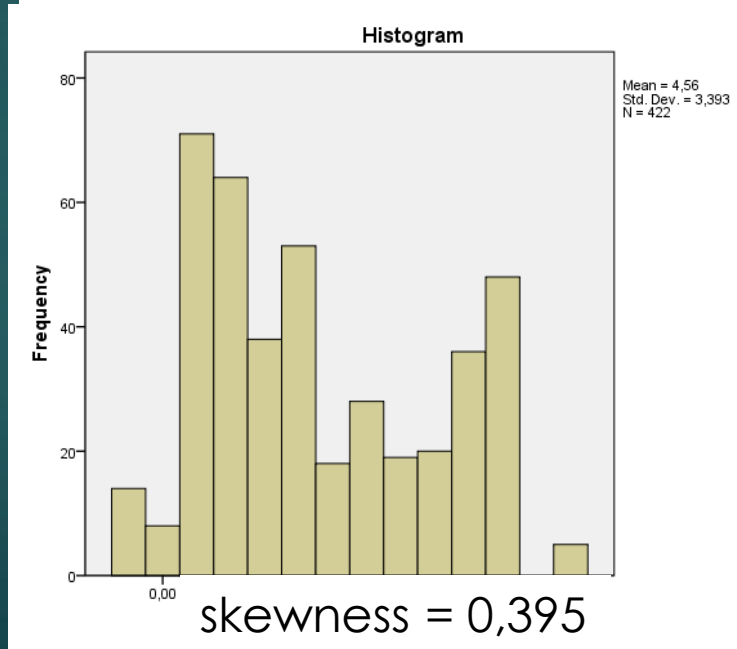


Verificação da Distribuição – Normal Q-Q Plot

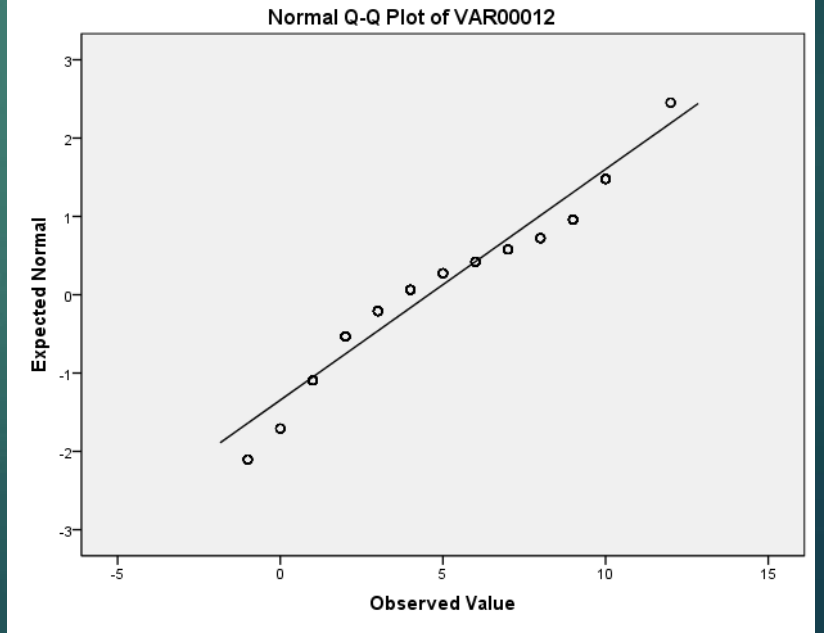


Skewness=0

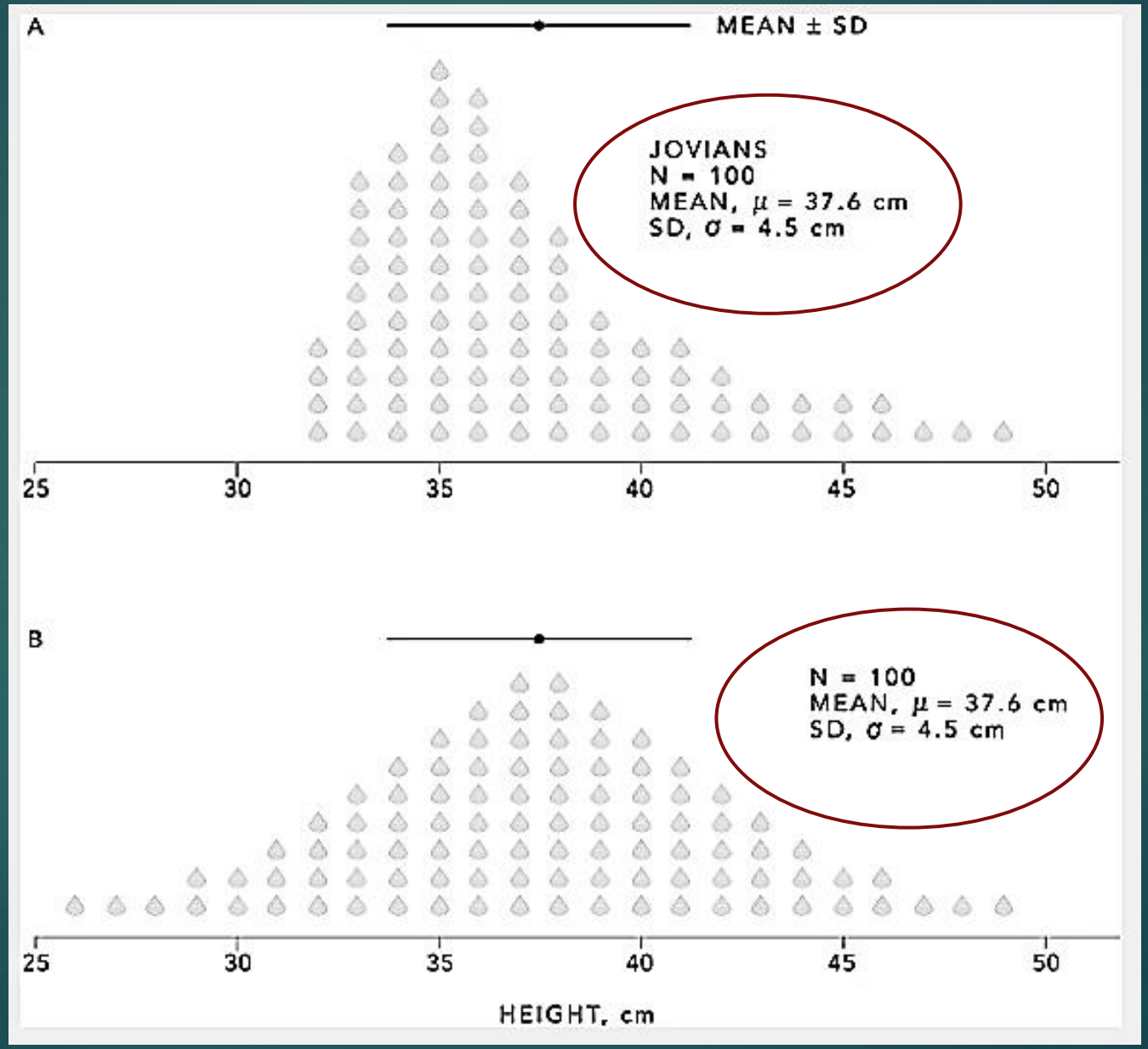




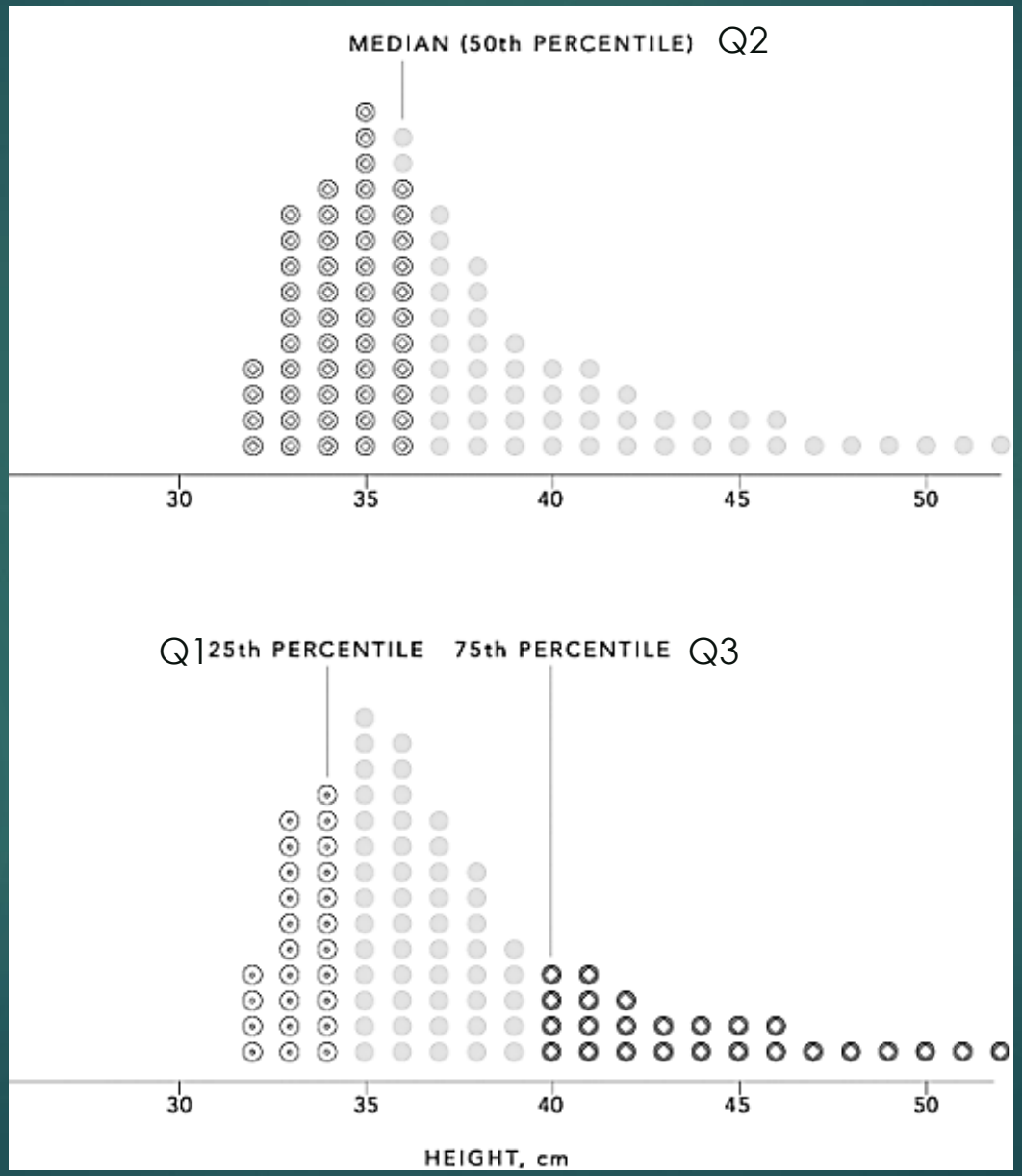
skewness = 0,395



DP na distribuição gaussiana x não-gaussiana

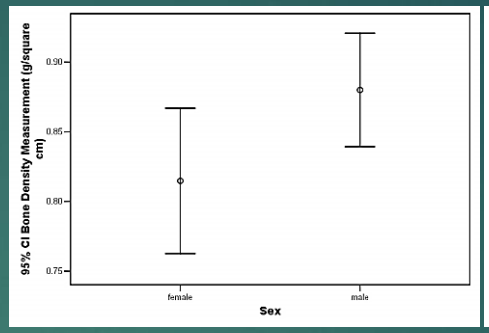


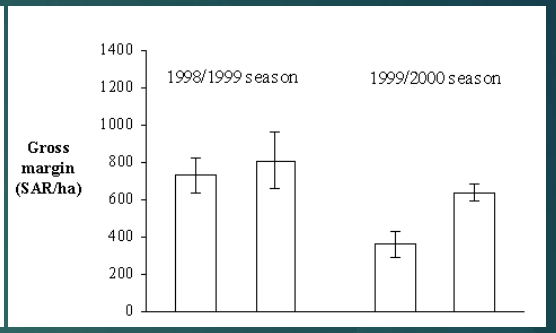
Mediana e Quartil



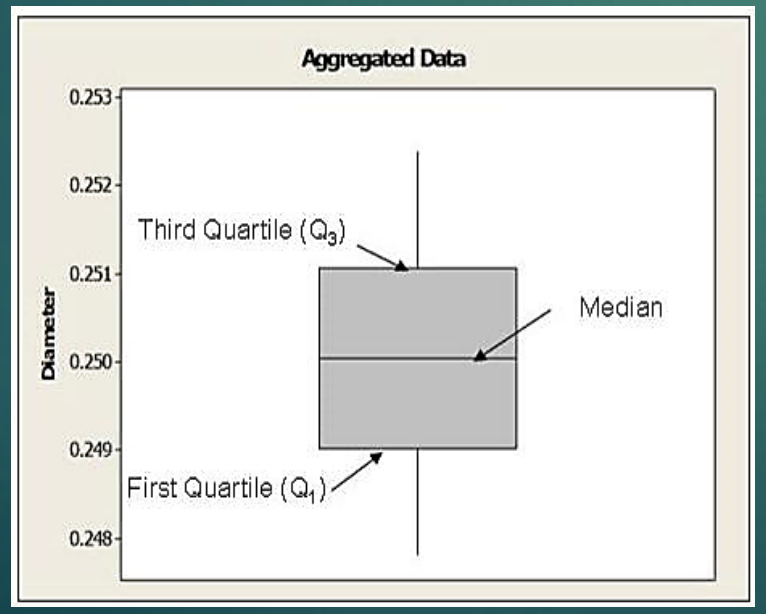
Resumindo

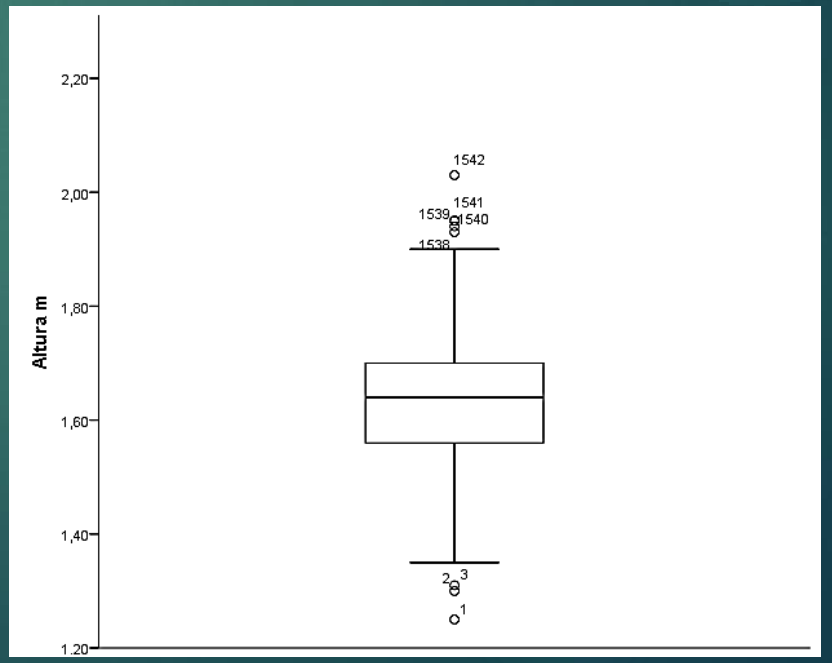
- Distribuição Normal
- Média ± Desvio-Padrão



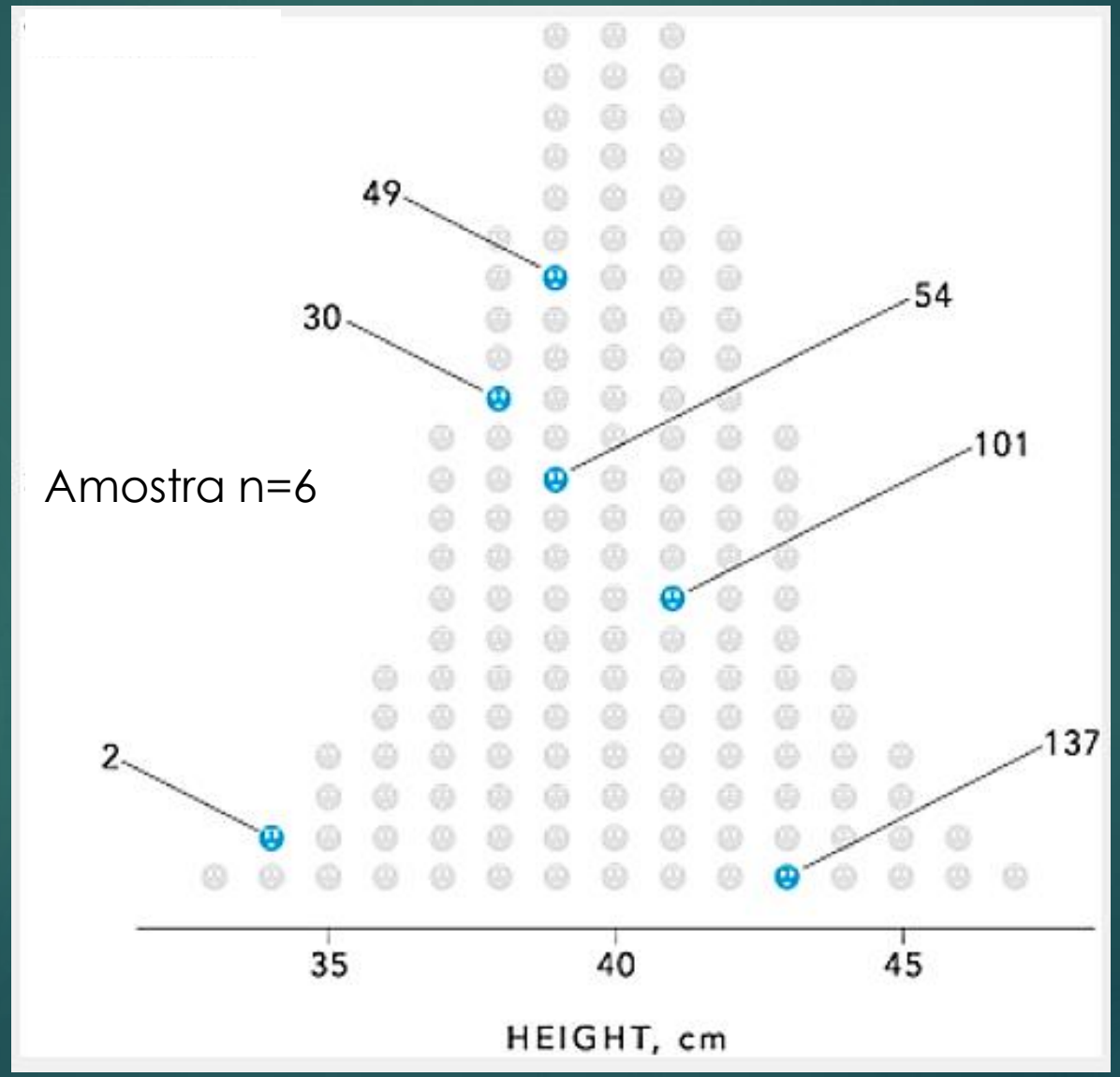


- Distribuição assimétrica (Não Normal)
  - Mediana e Q1 (25%) e Q3 (75%)





POPULAÇÃO X AMOSTRAS



Amostra n=6

- Média População e Desvio-Padrão população

$$\mu = \frac{\sum X}{N}$$

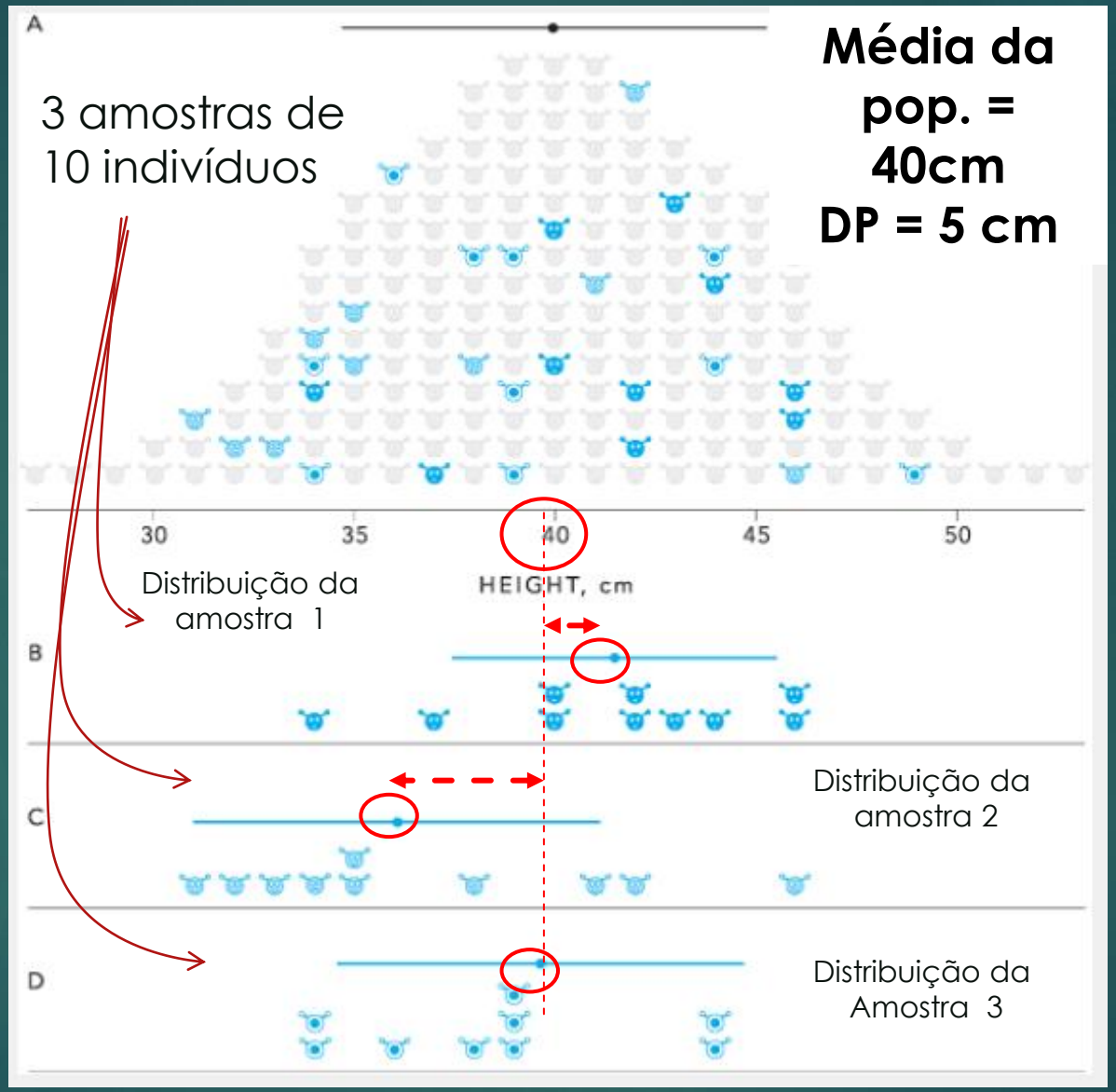
$$\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{N}}$$

- Média da amostra e desvio-padrão amostra

$$\bar{x} = \frac{\sum x}{n}$$

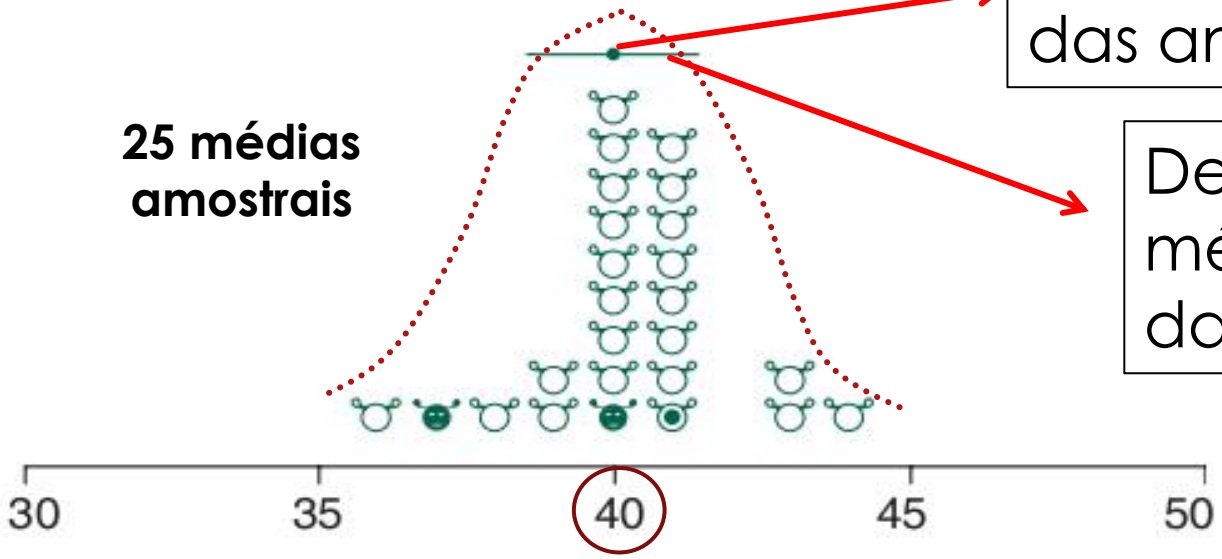
$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

O quão bem a **média da amostra** estima a **média da população?**


A
3 amostras de
10 indivíduos
Média da pop. = 40cm
DP = 5 cm
30
35
40
45
50
HEIGHT, cm
Distribuição da amostra 1
B
C
Distribuição da amostra 2
D
Distribuição da Amostra 3

Erro-padrão

- Tomemos 25 amostras de 10 indivíduos (pop.= 200)
- Façamos a distribuição das 25 médias amostrais
- Calculemos a média das médias e seu DP



Média das medias das amostras

Desvio-padrão da médias das médias das amostras

ERRO-PADRÃO

Média das amostras é semelhante a média da população, mas o DP das médias amostrais (erro-padrão) sempre será menor que DP da população.

ERRO-PADRÃO DA MÉDIA

- Quantifica a certeza com que a média de uma amostra aleatória estima a **Verdadeira Média da População** da qual a amostra foi retirada.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

Erro-padrão de amostras de uma população dada o desvio-padrão da população

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Erro-padrão da amostra com tamanho "n" e desvio-padrão "s"

Erro-padrão x Desvio-padrão

Medem coisas diferentes

DP: variabilidade na população

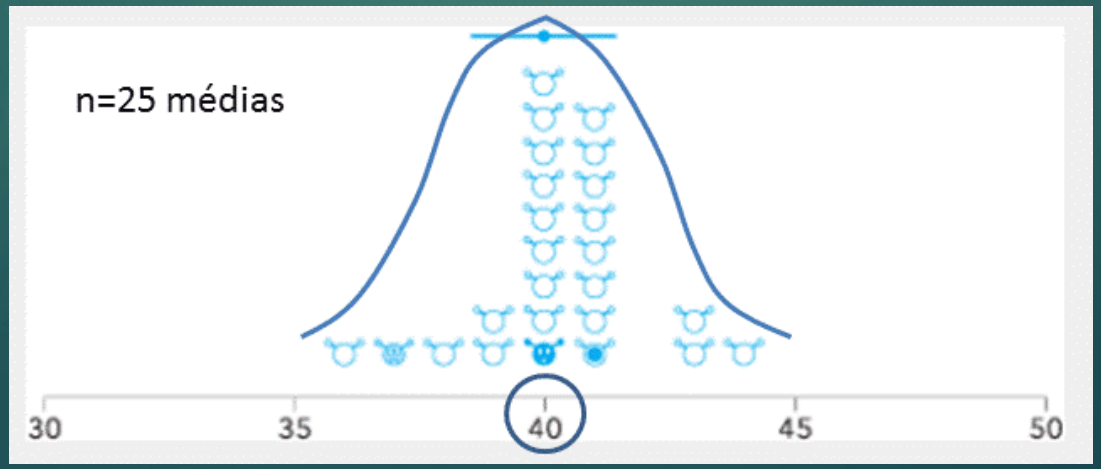
EP: incerteza na estimative da media populacional

Como as médias de todas as amostras aproximadamente segue uma distribuição normal a verdadeira média populacional (não observada) estará dentro de **± 2 erros-padrões da média das médias em 95% das vezes.**

Teorema do Limite Central

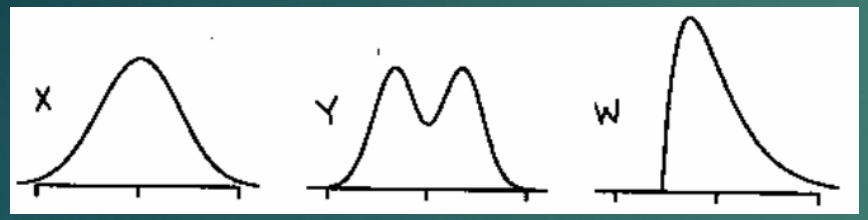
- **À medida que o tamanho ou número das amostras da mesma população aumentam, a distribuição das médias amostrais tende a uma distribuição normal**
- A média das médias amostrais será próximo a média populacional
- O desvio padrão das médias amostrais será o erro-padrão

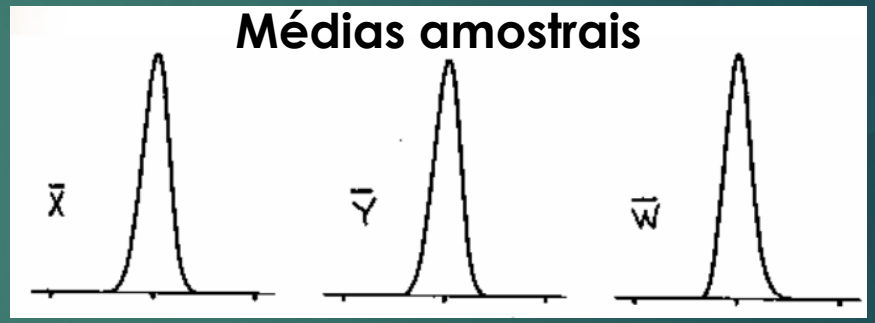
Distribuição gaussiana



Teorema do Limite Central

- ELE NOS DIZ QUE QUALQUER QUE SEJA A FORMA DA DISTRIBUIÇÃO ORIGINAL (POPULAÇÃO), A DISTRIBUIÇÃO DAS MÉDIAS AMOSTRAIS RESULTARÁ NUMA DISTRIBUIÇÃO NORMAL.

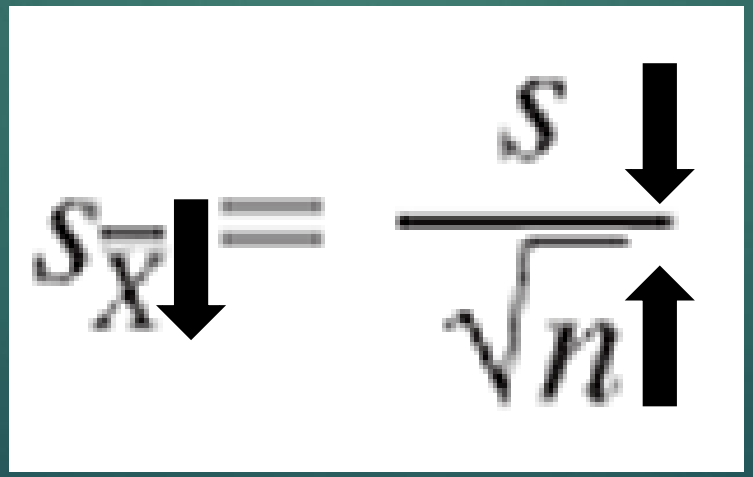




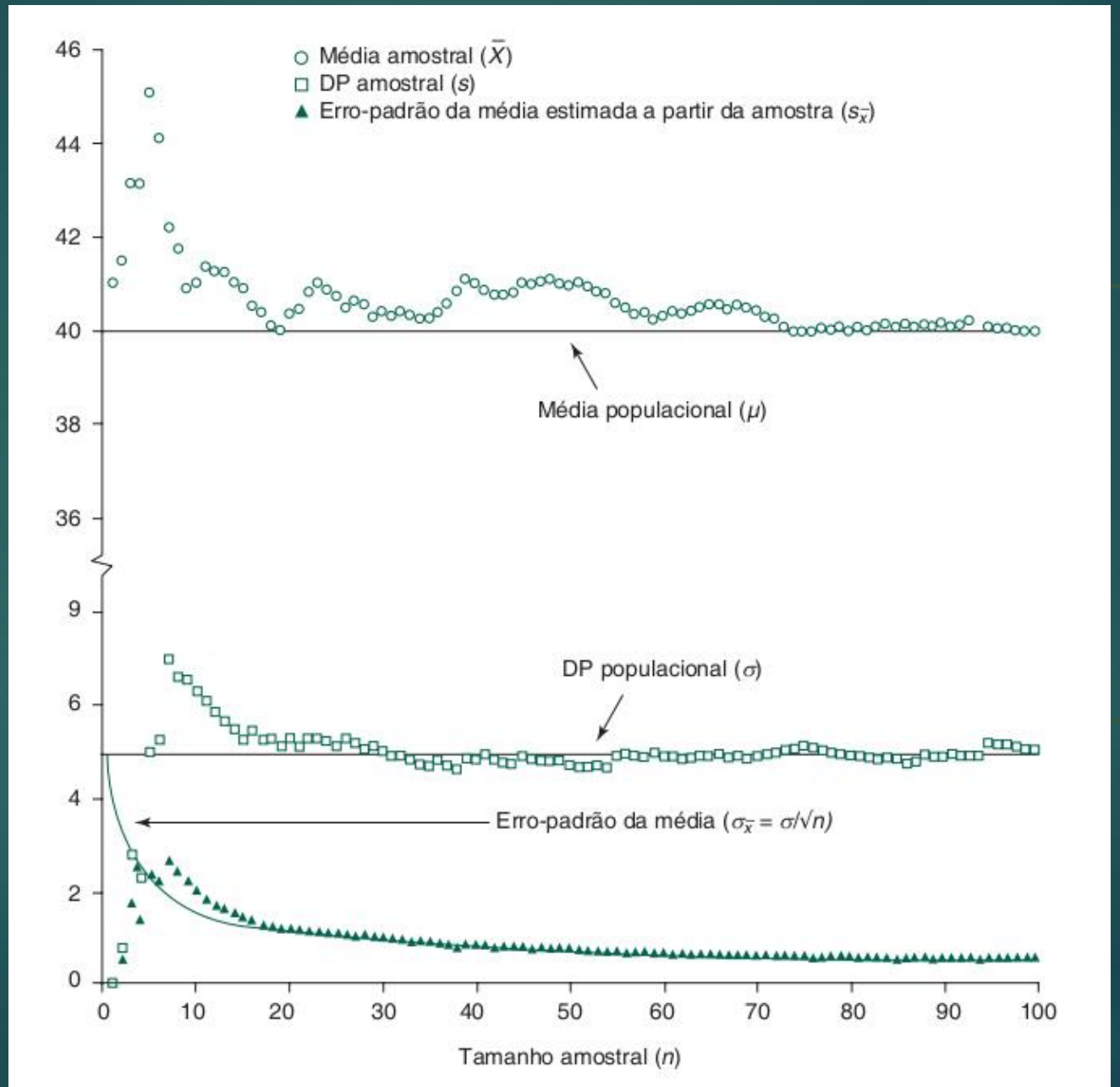
- A aproximação para a "normal" melhora à medida que o tamanho amostral cresce. Este resultado é conhecido como o Teorema Central do Limite e é notável porque permite-nos conduzir alguns procedimentos de inferência sem qualquer conhecimento da distribuição da população.

Resumo sobre Erro-Padrão

- **Quanto menor a variabiliade, menor o erro-padrão,**
- **Quanto maior a amostra (n), menor o erro-padrão**



Efeito do "n" na estimativa do DP, EP e média



Efeito do tamanho da amostra (n) na média e erro-padrão

Idade

100%

N	Mean	Std. Deviation
1421	56,03	14,917

Case Summaries

Idade

N	Mean	Std. Deviation	Std. Error of Mean
75	55,29	15,688	1,812

Case Summaries

Idade

N	Mean	Std. Deviation	Std. Error of Mean
419	55,81	14,849	,725

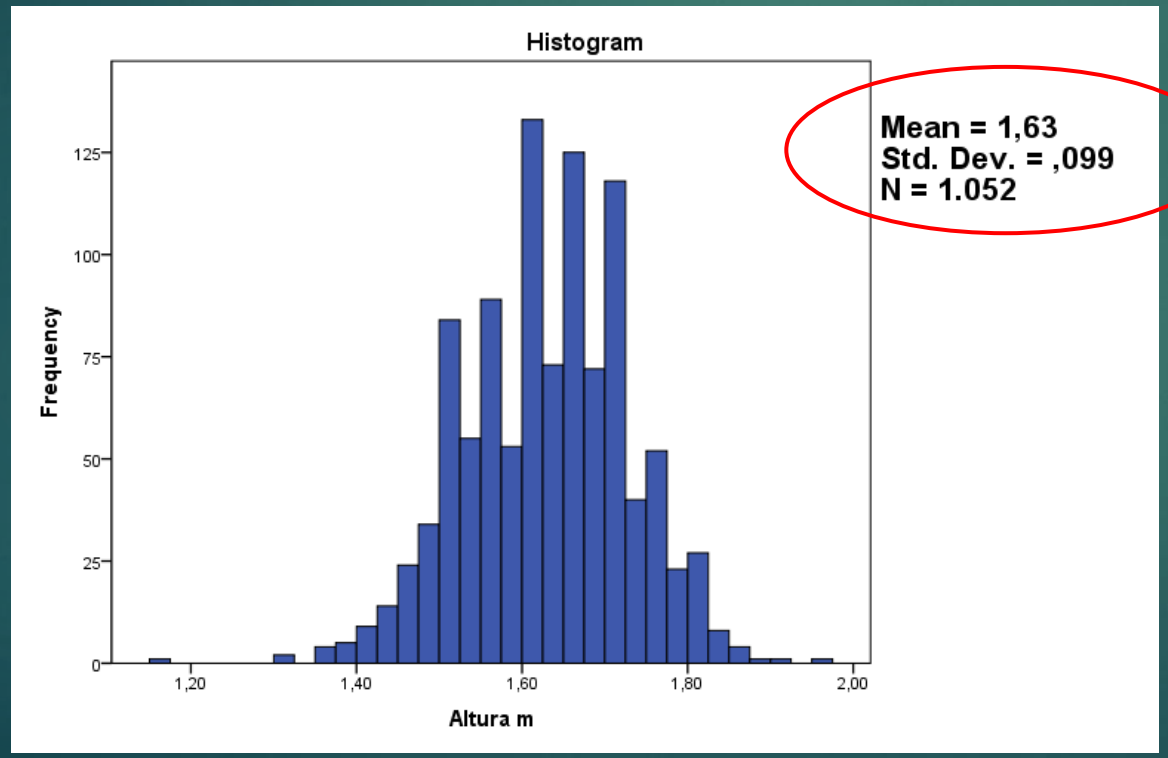
Case Summaries

Idade

N	Mean	Std. Deviation	Std. Error of Mean
850	56,46	14,882	,510

Intervalo de confiança (IC) da média

O intervalo de confiança de uma média nos fornece o "grau" de certeza (90%, 95%, 99%) de que o intervalo CONTÉM a VERDADEIRA MÉDIA POPULACIONAL

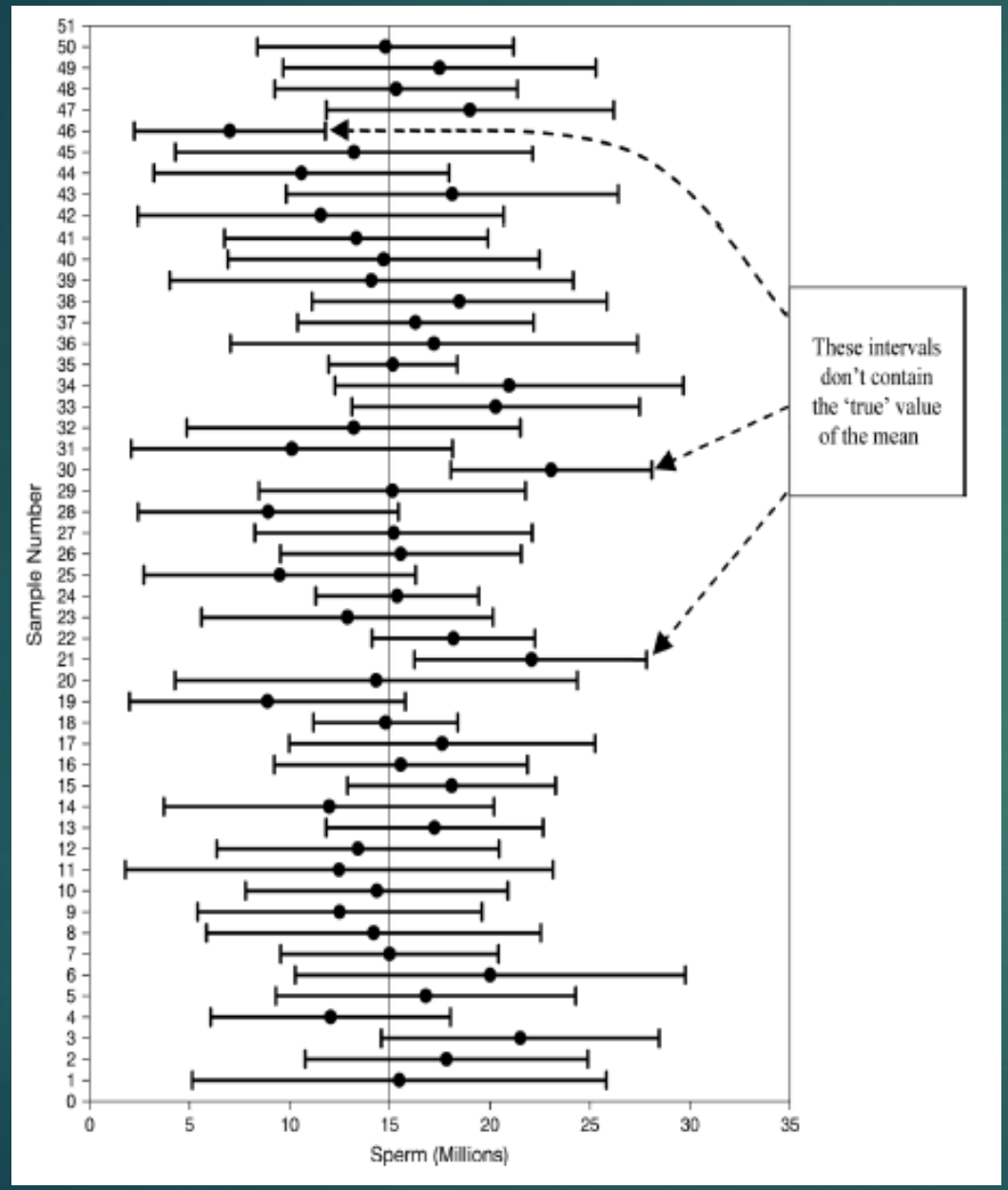


POPULAÇÃO

N= 50 (~5% da população)
média = 1,64
95% IC: 1,60 – 1,67

N= 200 (~20% da população)
média = 1,63
95% IC: 1,61 – 1,64

Intervalo de confiança (IC) da média



Se coletarmos "n" amostras e calcularmos a média e o IC95% destas "n" amostras, em 95% delas o intervalo de confiança conterá a verdadeira média populacional .

Intervalo de confiança (IC)

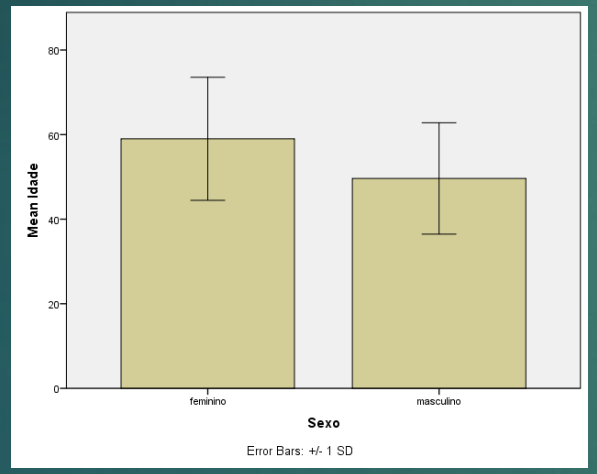
Condições necessárias para interpretação correta do IC

- Amostra deve ser aleatoriamente selecionada da população
- A distribuição da população é "normal"
- Todos os indivíduos são da mesma população e selecionados de forma independente

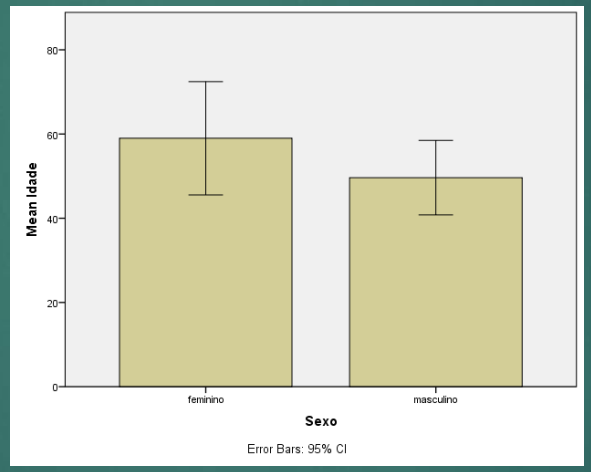
desvio-padrão x Erro-padão

N=1300

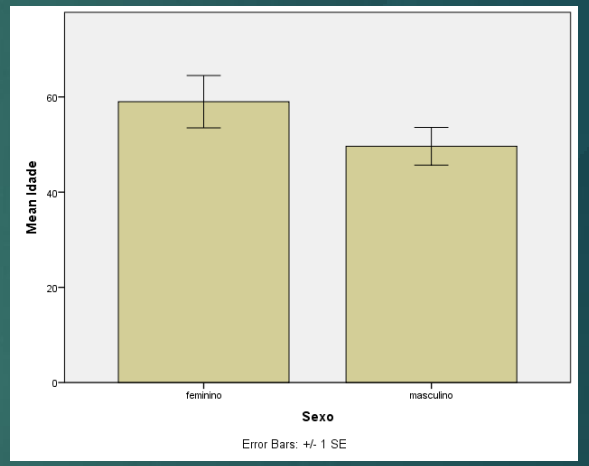
Média e DP



Média e IC 95%



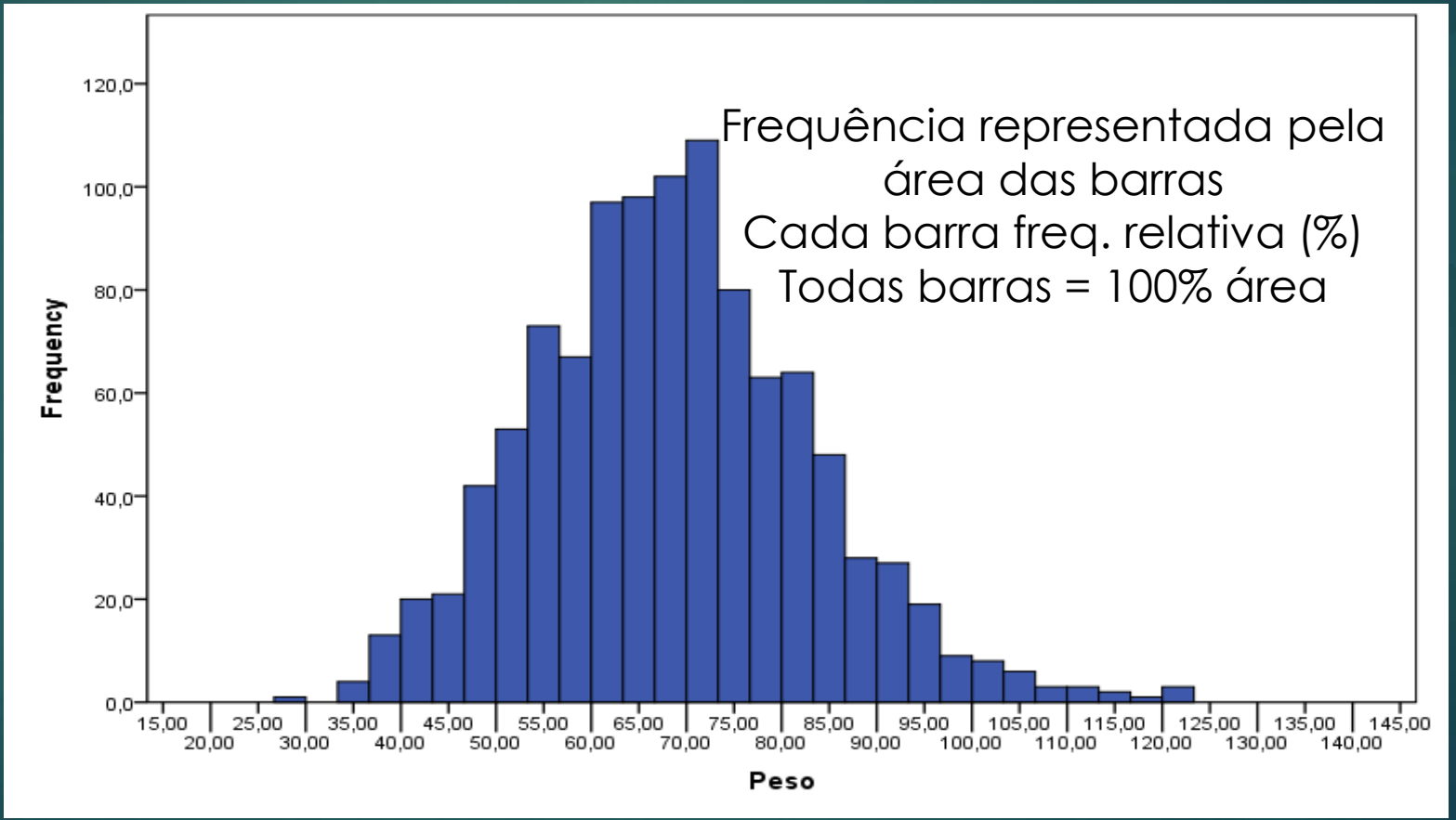
Média e erro-padrão



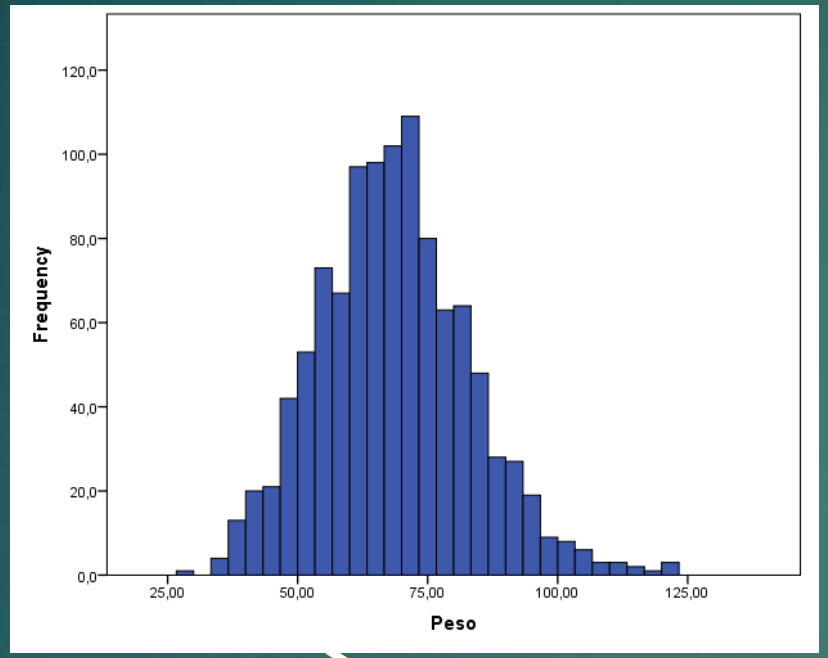
Representação gráfica para distribuições normais

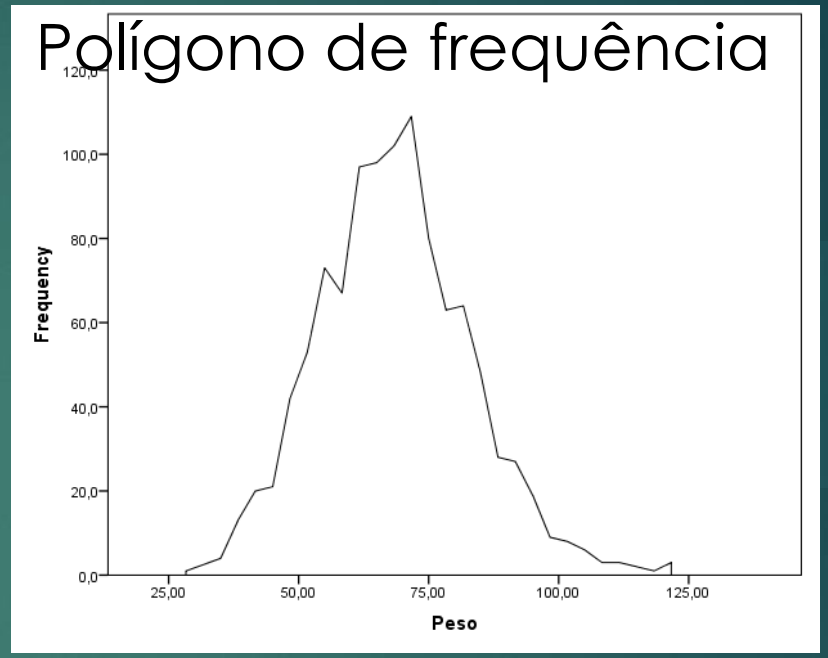
Histograma

Distribuição de frequência para dados contínuos ou discretos



Distribuiçao na População ou amostra





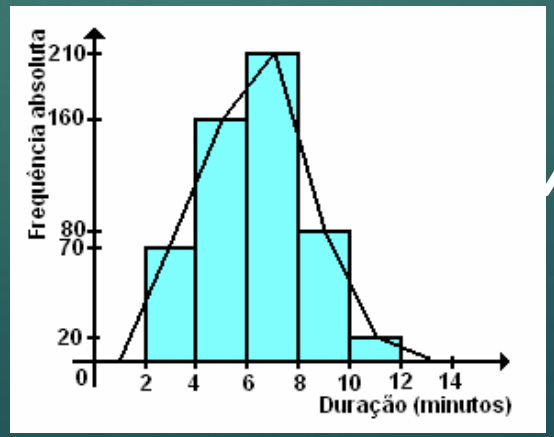
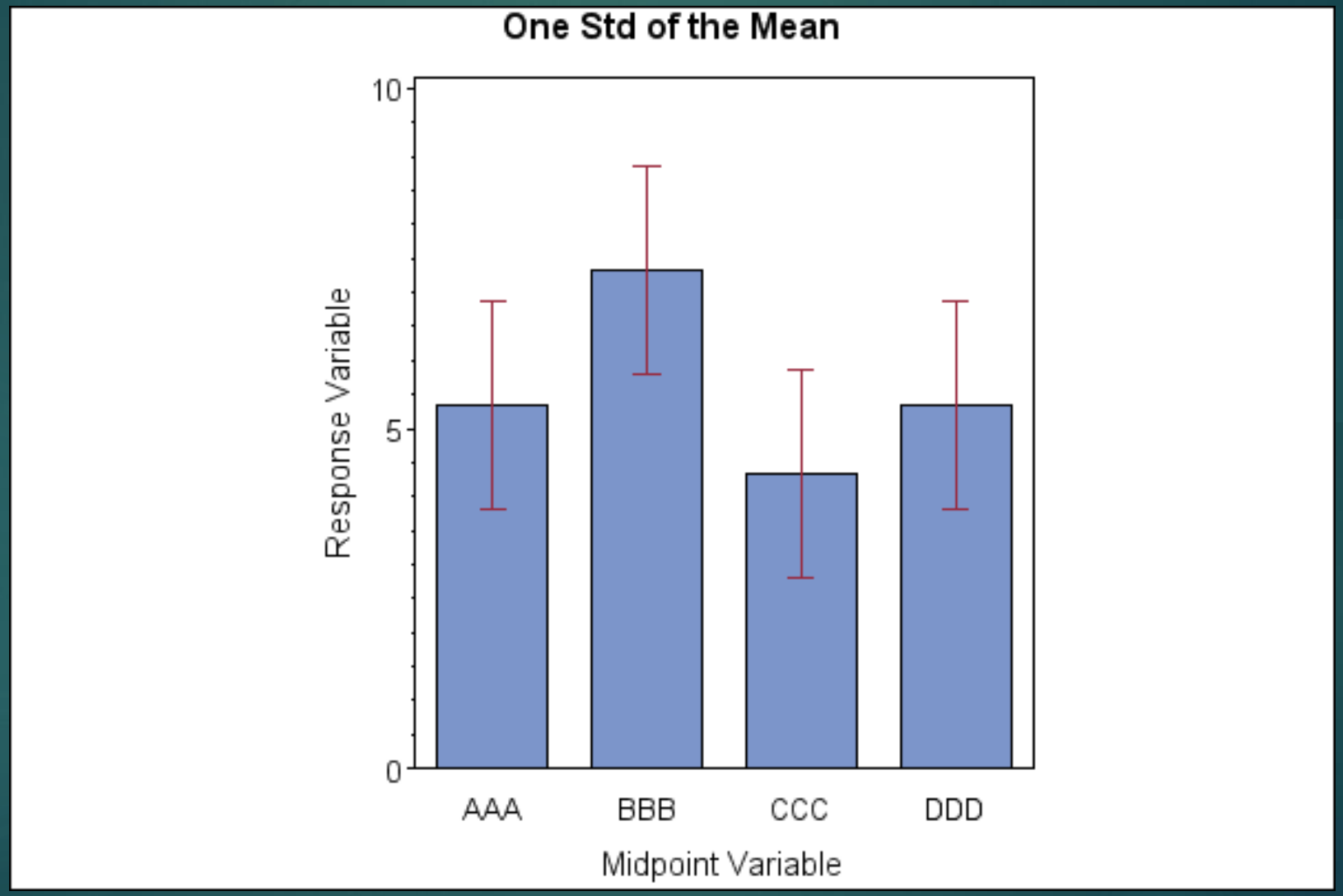


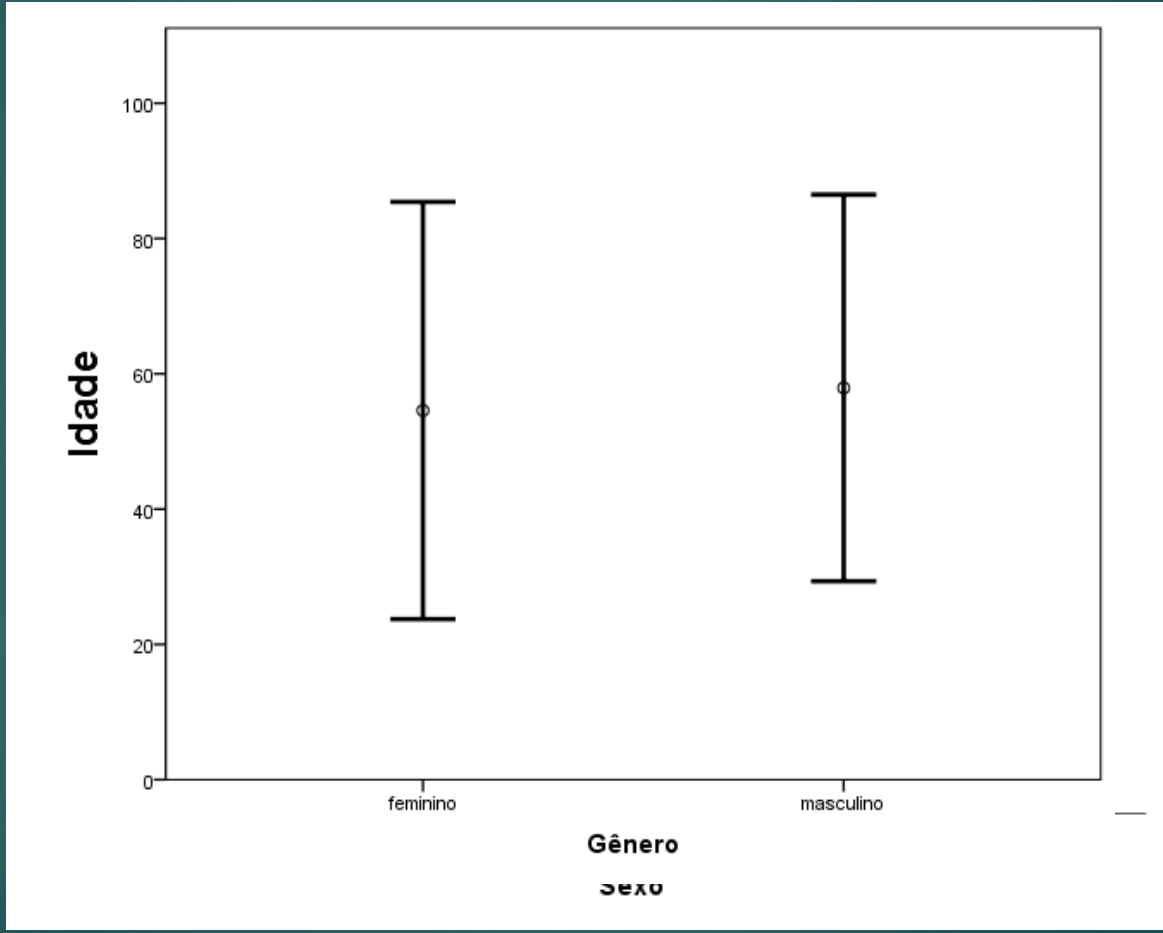
Gráfico de Barras com DP



Gráficos – Error Bars

Variáveis quantitativas

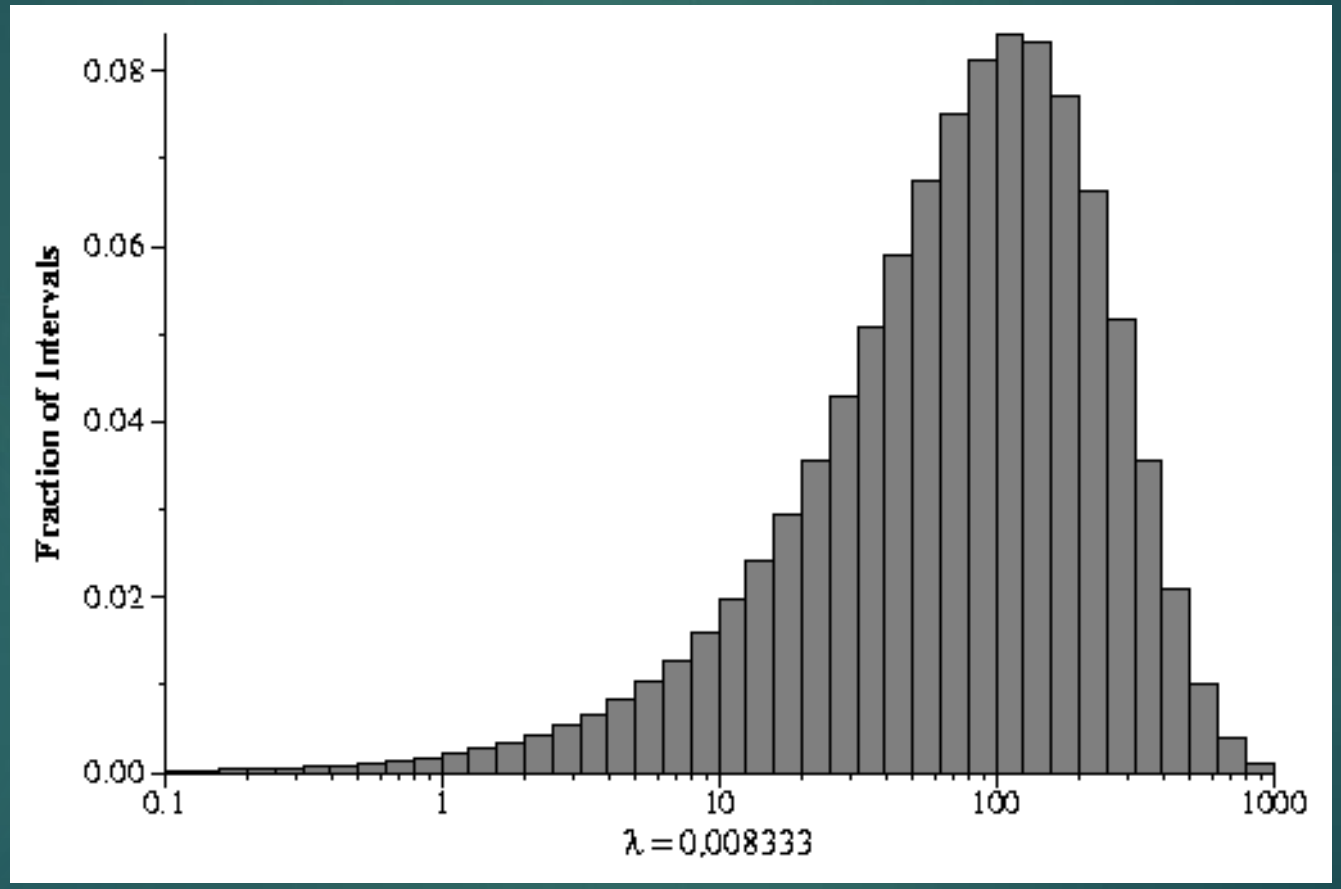
Medidas Centrais com dispersão ou variabilidade



Útil apenas nas distribuições normais

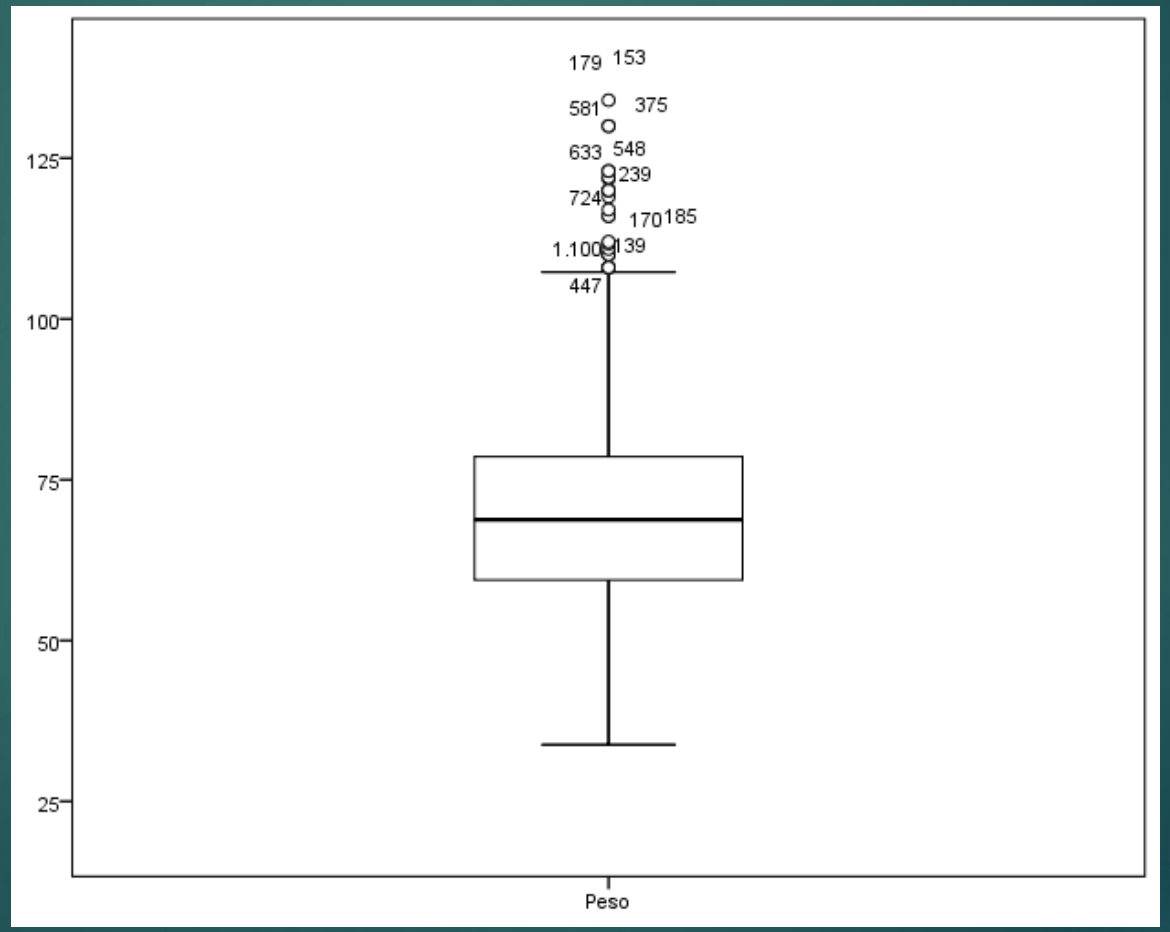
Representação gráfica para distribuições assimétricas

Distribuição assimétrica

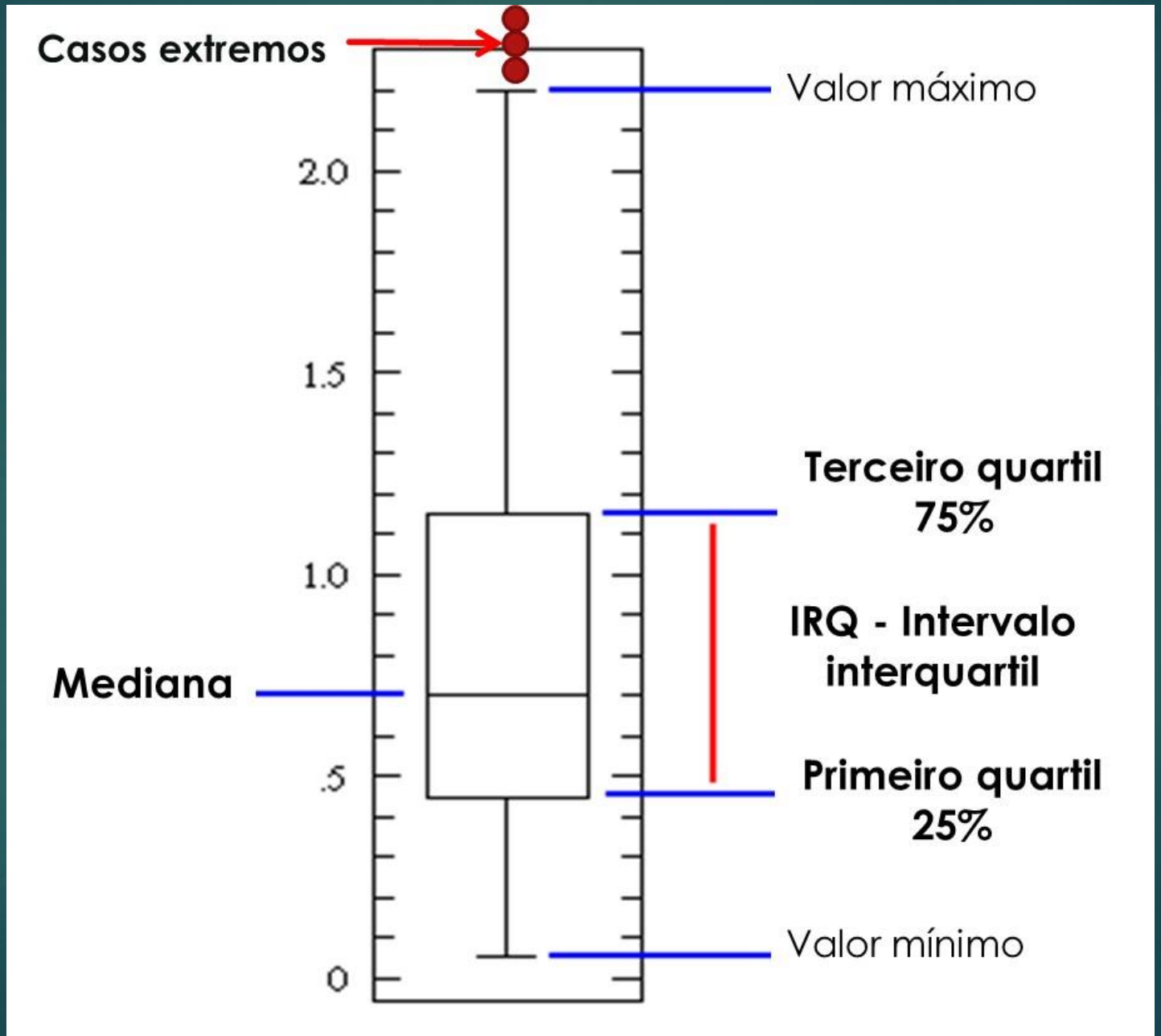


Box and Whisker Plot

Preferência para distribuição assimétrica



Box Plot


Casos extremos
Valor máximo
2.0
1.5
Terceiro quartil
75%
1.0
IRQ - Intervalo
interquartil
Mediana
.5
Primeiro quartil
25%
Valor mínimo
0

Box Plot
valores externos e extremos