

AULA 01

Estatísticas Descritivas

Ernesto F. L. Amaral

13 de agosto de 2012

Faculdade de Filosofia e Ciências Humanas (FAFICH)
Universidade Federal de Minas Gerais (UFMG)

Fonte:

Babbie, Earl. 1999. “Métodos de Pesquisas de Survey.” Belo Horizonte: Editora UFMG. Capítulo 4 (pp.93-111).

Triola, Mario F. 2008. “Introdução à estatística”. 10^a ed. Rio de Janeiro: LTC. Capítulo 1 (pp.2-31) e Capítulo 3 (pp.60-109).

Wooldridge, Jeffrey M. 2008. “Introdução à econometria: uma abordagem moderna.” São Paulo: Cengage Learning. Capítulo 1 (pp.1-17).

ESQUEMA DA AULA

- Tipos de dados.
- Bancos de dados.
- Medidas de centro.
- Medidas de variação.
- Medidas de posição relativa.
- Análise exploratória de dados (AED).

ESTATÍSTICA DESCRITIVA E INFERÊNCIA ESTATÍSTICA

- Triola afirma que estatística descritiva e inferência estatística são as duas divisões gerais do objeto da estatística.
- Neste momento, estamos trabalhando com métodos de estatística descritiva, já que objetivo é de resumir ou descrever as características importantes de um conjunto de dados.
- Ao final do curso, usaremos métodos de inferência estatística (regressão), com objetivo de fazer generalizações sobre uma população, utilizando dados amostrais.
- Ou seja, a inferência estatística visa realizar análises que vão além dos dados conhecidos.

TIPOS DE DADOS

TIPOS DE DADOS

- **Dados** são observações coletadas de um determinado grupo de interesse.
- **Dados quantitativos** são números que representam contagens ou medidas (renda, anos de escolaridade...).
 - **Dados discretos** são aqueles em que o número de valores possíveis são finitos ou “enumeráveis” (número de cômodos em um domicílio...).
 - **Dados contínuos** resultam de infinitos valores possíveis em uma escala contínua (renda per capita...).
- **Dados qualitativos** (ou categóricos ou de atributos) podem ser separados em diferentes categorias que se distinguem por alguma característica não-numérica (sexo, ideologia política).

NÍVEIS DE MENSURAÇÃO DE DADOS

- Nível **nominal** de mensuração possui dados que informam nomes, rótulos ou categorias:
 - Os dados não são ordenados e não devem ser usados para cálculos de médias.
 - Raça e código postal, por exemplo.
- Nível **ordinal** de mensuração engloba dados que podem ser organizados em alguma ordem:
 - Sabemos que há diferenças relativas entre os valores dos dados, mas não sabemos as magnitudes das diferenças.
 - Na escala de frequência (pouco/médio/muito), é possível ordenar os dados, mas não sabemos se a diferença entre “pouco” e “médio” é o mesmo que “médio” e “muito”.

NÍVEIS DE MENSURAÇÃO DE DADOS (cont.)

- Nível **intervalar** de mensuração é similar ao ordinal, mas sabemos as magnitudes das diferenças entre dois valores:
 - Os dados não possuem um ponto inicial zero natural.
 - Sabemos as magnitudes das diferenças entre os anos censitários (1970, 1980, 1991 e 2000), mas o tempo não começou em zero.
- Nível de mensuração de **razão** é similar ao intervalar, mas há um ponto inicial zero natural:
 - Como há um zero que indica nenhuma quantidade, é possível dizer que uma quantidade é maior que outra em X vezes (razões significativas).
 - 30 anos de idade é 6 vezes maior do que 5 anos de idade, por exemplo.

RESUMO DOS NÍVEIS DE MENSURAÇÃO DE DADOS

Nível	Resumo	Exemplo
Nominal	Apenas categorias. Os dados não podem ser arranjados em um esquema de ordem. Há categorias ou nomes apenas.	Município de residência.
Ordinal	As categorias são ordenadas, mas as diferenças não podem ser encontradas ou não têm significado.	Frequência à igreja: pouco, médio, muito.
Intervalar	As diferenças são significativas, mas não existe ponto inicial zero natural e as razões não têm sentido.	Ano censitário (não há tempo zero).
Razão	Há um ponto inicial zero natural e as razões são significativas.	Taxa de desemprego.

TIPOS DE BANCO DE DADOS

DESENHOS BÁSICOS DE *SURVEY*

- Após especificar os objetivos e unidades de análise da pesquisa, é preciso escolher entre diversos desenhos diferentes (Babbie, 1999):
 - *Surveys* interseccionais (*cross-sectional*).
 - *Surveys* longitudinais (tendências, coortes ou painel).
 - *Surveys* interseccionais servindo como longitudinais.
- Wooldridge (2008) classifica os dados econômicos em:
 - Dados de corte transversal = *surveys* interseccionais.
 - Cortes transversais agrupados = estudos de tendências.
 - Dados de séries de tempo = estudos de coortes.
 - Dados de painel ou longitudinais = estudos de painel.

DADOS DE CORTE TRANSVERSAL (Wooldridge)

SURVEYS INTERSECCIONAIS (Babbie)

- Um conjunto de dados de corte transversal consiste em uma amostra de uma unidade de análise, tomada em um determinado ponto no tempo.
- Esses dados são muito utilizados em economia e em outras ciências sociais.
- Dados em um determinado ponto do tempo são importantes para testar hipóteses e avaliar políticas.
- Dados podem ter problemas de seleção amostral, no caso de determinados indivíduos não revelarem informações acuradas.
- Amostragem deve ser realizada de forma acurada para evitar que coleta se concentre em unidades com características semelhantes.

EXEMPLO DE DADOS DE CORTE TRANSVERSAL

– Conjunto de dados de corte transversal para o ano de 1976 de 526 trabalhadores (Wooldridge 2008):

Número da observação	Salário por hora	Anos de escolaridade	Anos de experiência no mercado de trabalho	Feminino	Estado civil (casado)
1	3,10	11	2	1	0
2	3,24	12	22	1	1
3	3,00	11	2	0	0
4	6,00	8	44	0	1
5	5,30	12	7	0	1
...
525	11,56	16	5	0	1
526	3,50	14	5	1	0

CORTES TRANSVERSAIS AGRUPADOS (Wooldridge)

ESTUDOS DE TENDÊNCIAS (Babbie)

- Uma população pode ser amostrada e estudada em ocasiões diferentes.
- Um mesmo conjunto de variáveis é coletado em diferentes períodos do tempo, em **distintas** amostras aleatórias de uma mesma população (Censo Demográfico, Pesquisa Nacional por Amostra de Domicílios – PNAD).
- Agrupar cortes transversais de diferentes anos é eficaz para analisar os efeitos de uma política pública.
- O ideal é coletar dados de anos anteriores e posteriores a uma importante mudança de política governamental.
- Além de aumentar o tamanho da amostra, a análise de corte transversal agrupada é importante para estimar como uma relação fundamental mudou ao longo do tempo.
- Geralmente são utilizados dados secundários, coletados por outros pesquisadores ou instituições.

EXEMPLO DE CORTES TRANSVERSAIS AGRUPADOS

– Conjunto de dados sobre os preços da moradia em 1993 e 1995 nos Estados Unidos (Wooldridge 2008):

Número da observação	Ano	Preço comercializado	Impro	Arquad	Quantidade de dormitórios	Quantidade de banheiros
1	1993	85.500	42	1.600	3	2,0
2	1993	67.300	36	1.440	3	2,5
3	1993	134.000	38	2.000	4	2,5
...
250	1993	243.600	41	2.600	4	3,0
251	1995	65.000	16	1.250	2	1,0
252	1995	182.400	20	2.200	4	2,0
253	1995	97.500	15	1.540	3	2,0
...
520	1995	57.200	16	1.100	2	1,5

DADOS DE SÉRIES DE TEMPO (Wooldridge)

ESTUDOS DE COORTES (Babbie)

- Um conjunto de dados de séries de tempo consiste em observações sobre variáveis ao longo do tempo.
- Como eventos passados podem influenciar eventos futuros, o tempo é uma dimensão importante em um conjunto de dados de séries de tempo.
- A análise desses dados pode ser dificultada, porque observações econômicas não são independentes ao longo do tempo (variáveis possuem padrões sazonais).
- Há uma série de frequências possíveis: diárias, semanais, mensais, trimestrais, anuais, decenais...
- Estes dados são também chamados de estudos de coorte, em que mesma população é analisada, mas amostras estudadas podem ser diferentes:
 - Pessoas com 10 anos em 2000, 20 anos em 2010, 30 anos em 2020, 40 anos em 2030...

EXEMPLO DE DADOS DE SÉRIES DE TEMPO

– Conjunto de dados de séries de tempo sobre efeitos do salário mínimo em Porto Rico (apud Wooldridge 2008):

Número da observação	Ano	Salário mínimo médio no ano	Taxa de trabalhadores cobertos pela lei de salário mínimo	Taxa de desemprego	Produto Nacional Bruto (PNB)
1	1950	0,20	20,1	15,4	878,7
2	1951	0,21	20,7	16,0	925,0
3	1952	0,23	22,6	14,8	1.015,9
...
37	1986	3,35	58,1	18,9	4.281,6
38	1987	3,35	58,2	16,8	4.496,7

DADOS DE PAINEL OU LONGITUDINAIS (Wooldridge) ESTUDOS DE PAINEL (Babbie)

- Um conjunto de dados de painel consiste em uma série de tempo para **cada** membro do corte transversal.
- Os dados de painel são distintos dos dados de corte transversal agrupados (tendências) e de séries de tempo (coortes), porque as **mesmas** unidades são acompanhadas ao longo de um determinado período.
- Dados de painel podem ser coletados para indivíduos, domicílios, instituições ou unidades geográficas.
- Esses dados são os mais sofisticados para fins explicativos, mas são mais difíceis e caros de se obter.
- Pode haver problema de grande número de não respostas nas últimas ondas de entrevistas.
- A análise dos dados pode se tornar complicada quando se tentar avaliar as mudanças dos indivíduos no tempo.

EXEMPLO DE DADOS DE PAINEL OU LONGITUDINAIS

- Conjunto de dados de painel sobre crime e estatísticas relacionadas em 1986 e 1990 em 150 cidades nos Estados Unidos (Wooldridge 2008):

Número da observação	Cidade	Ano	Homicídios	População	Desemprego	Polícia
1	1	1986	5	350.000	8,7	440
2	1	1990	8	359.200	7,2	471
3	2	1986	2	64.300	5,4	75
4	2	1990	1	65.100	5,5	75
...
297	149	1986	10	260.700	9,6	286
298	149	1990	6	245.000	9,8	334
299	150	1986	25	543.000	4,3	520
300	150	1990	32	546.200	5,2	493

MEDIDAS DE CENTRO

MEDIDAS DE CENTRO

- Medida de centro é um valor no centro ou meio do conjunto de dados.
- Desejamos obter um número que represente o valor central de um conjunto de dados.
- Os conceitos e métodos para encontrar média e mediana devem ser bem entendidos.
- O valor da média pode ser muito afetado pela presença de um valor discrepante (“outlier”), mas a mediana não é tão sensível a um “outlier”.

MÉDIA

- **Média aritmética** é calculada pela adição dos valores de uma variável e divisão deste total pelo número de valores.
- Essa medida é muito utilizada na descrição de dados.

$$Média = \frac{\sum x}{n}$$

- **Estatísticas amostrais** são usualmente representadas por letras do alfabeto latino e minúsculas:

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

- **Parâmetros populacionais** são representados por letras gregas e maiúsculas:

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

MEDIANA

– **Mediana** é o valor do meio quando os dados originais estão organizados em ordem crescente (ou decrescente) de magnitude (\tilde{x}).

– Para **encontrar a mediana**:

1) Ordene os valores de uma variável.

2) Se o número de valores for ímpar, a mediana será o número localizado no meio exato da lista.

ou

2) Se o número de valores for par, a mediana será encontrada pelo cálculo da média dos dois números do meio.

– A média é afetada por **valores extremos**, ao contrário da mediana. Por isso, quando temos “outliers”, mediana pode ser mais apropriada.

MODA

- A **moda** de um conjunto de dados é o valor que ocorre com maior frequência.
- Conjunto de dados **bimodal**: quando dois valores ocorrem com maior frequência, cada um é uma moda.
- Conjunto de dados **multimodal**: quando mais de dois valores ocorrem com maior frequência.
- Quando nenhum valor se repete, não há moda.
- Moda não é muito usada com dados numéricos.
- Dentre as medidas de centro consideradas, é a única que pode ser usada com dados no nível nominal de mensuração (nomes, rótulos e categorias).
- Não faz muito sentido realizar cálculos numéricos (média e mediana) com dados categóricos.

PONTO MÉDIO

- **Ponto médio** é a medida de centro que é exatamente o valor a meio caminho entre o maior valor e o menor valor no conjunto original de dados.
- É encontrado pela soma do maior valor e o menor valor dos dados, dividindo-se a soma por 2:

$$\text{ponto médio} = \frac{\text{valor máximo} + \text{valor mínimo}}{2}$$

- É raramente utilizado, já que é muito sensível a valores extremos.
- Vantagens: (1) fácil de calcular; e (2) evidencia que há diferentes maneiras de definir centro dos dados.
- Não deve ser confundido com mediana.

REGRA DE ARREDONDAMENTO

- Use uma casa decimal a mais do que é apresentado no conjunto original de valores:
 - A média de 80,4 e 80,6 é igual a 80,50.
- Quando valores originais são números inteiros, arredondamos para o décimo mais próximo:
 - A média de 2, 3, 5 é igual a 3,3.
- Arredonde apenas a resposta final e não os valores intermediários que surgirem durante os cálculos.

MÉDIA DE UMA DISTRIBUIÇÃO DE FREQUÊNCIA

- A média de uma população não é necessariamente igual à média das médias de diferentes subconjuntos da população.
- Quando usamos dados resumidos em uma distribuição de frequência, devemos considerar o ponto médio de cada classe, pois não temos os valores de cada observação.
- Por exemplo, o intervalo de classe de 21-30 (anos) assumirá o valor de 25,5 (ponto médio da classe).
- Procedimento:
 - 1) Multiplique cada frequência pelo ponto médio da classe e adicione os produtos: $\sum(f * x)$
 - 2) Adicione as frequências: $\sum f$
 - 3) Divida 1 por 2: $\sum(f * x) / \sum f$

EXEMPLO

Idade da atriz	Frequência (<i>f</i>)	Ponto médio da classe (<i>x</i>)	<i>f</i> * <i>x</i>
21-30	28	25,5	714
31-40	30	35,5	1.065
41-50	12	45,5	546
51-60	2	55,5	111
61-70	2	65,5	131
71-80	2	75,5	151
Total	76	---	2.718

$$\bar{x} = \frac{\sum(f * x)}{\sum f} = \frac{2.718}{76} = 35,8$$

MÉDIA PONDERADA

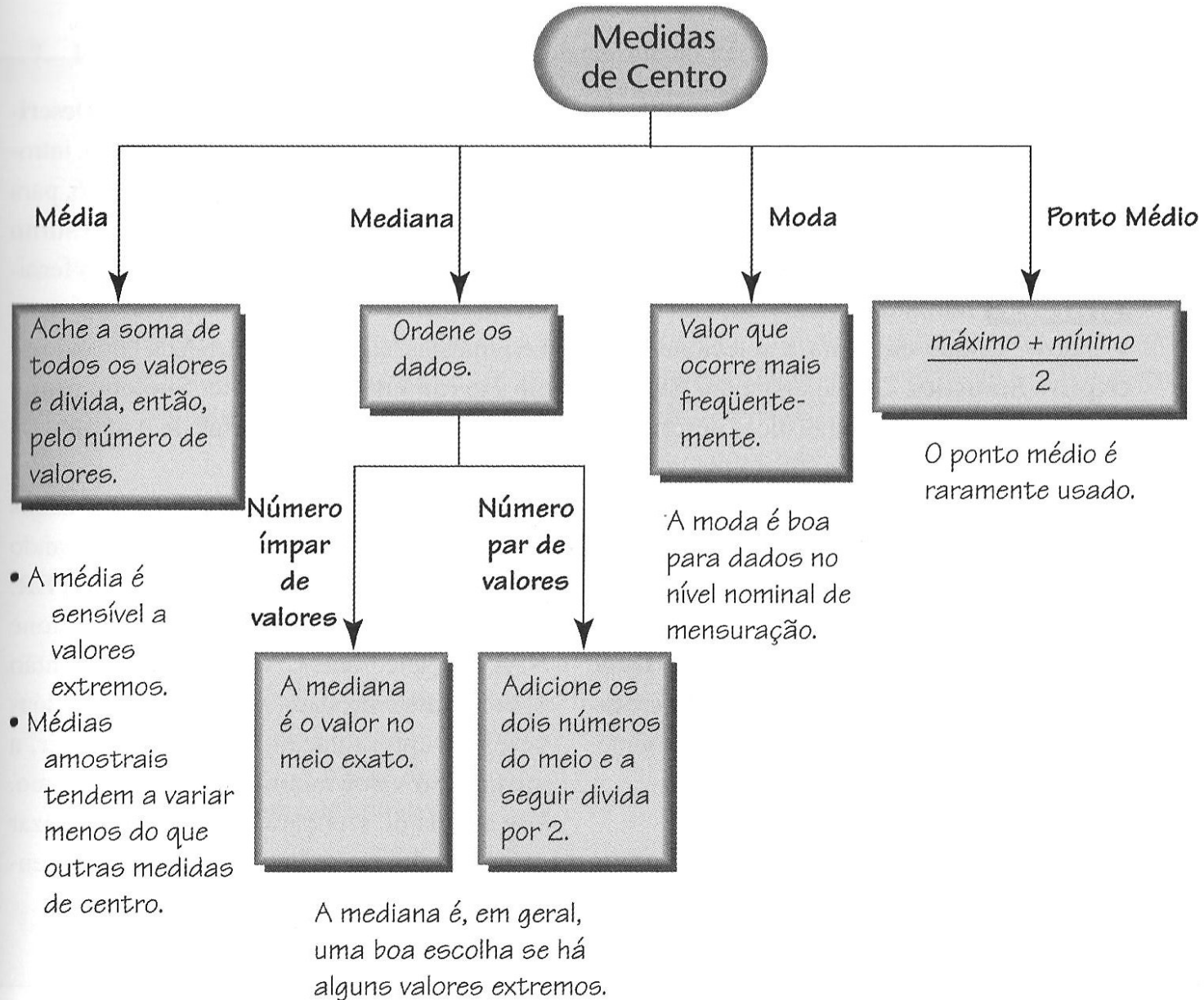
- Média ponderada dos valores de x é uma média calculada com os diferentes valores, associados a diferentes pesos (representados por w).

$$\bar{x} = \frac{\sum(w * x)}{\sum w}$$

- Por exemplo, suponha uma disciplina com três exercícios, valendo 30%, 30% e 40% da nota final.
- Suponha que um aluno recebeu as notas: 70, 85, 80.
- A nota final será:

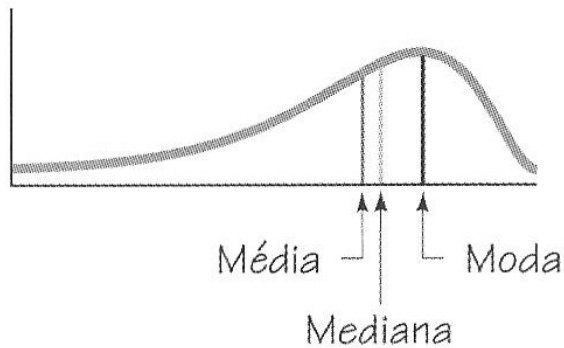
$$\bar{x} = \frac{(30 * 70) + (30 * 85) + (40 * 80)}{30 + 30 + 40} = \frac{7.850}{100} = 78,5$$

RESUMO DE MEDIDAS DE CENTRO

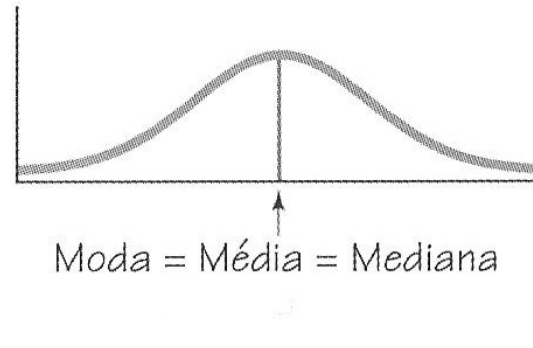


ASSIMETRIA

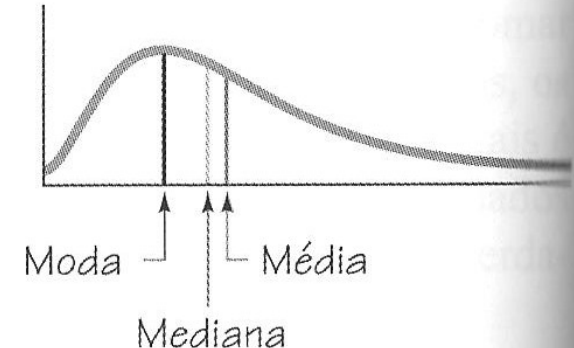
- Uma distribuição de dados é assimétrica quando se estende mais para um lado do que para o outro.
- A distribuição é simétrica se a metade esquerda de seu histograma é praticamente igual à sua metade direita.



(a) Assimétrica à Esquerda (Negativamente Assimétrica): A média e a mediana estão à esquerda da moda.



(b) Simétrica (Assimetria Zero): A média, mediana e moda são iguais.



(c) Assimétrica à Direita (Positivamente Assimétrica): A média e a mediana estão à direita da moda.

- Distribuições assimétricas à direita são mais comuns do que assimétricas à esquerda.

MEDIDAS DE VARIAÇÃO

MEDIDAS DE VARIAÇÃO

– Tempo médio de espera é igual nestas distribuições (6 min):

Banco 1: Filas de espera variáveis	6	6	6
Banco 2: Fila única de espera	4	7	7
Banco 3: Filas múltiplas de espera	1	3	14

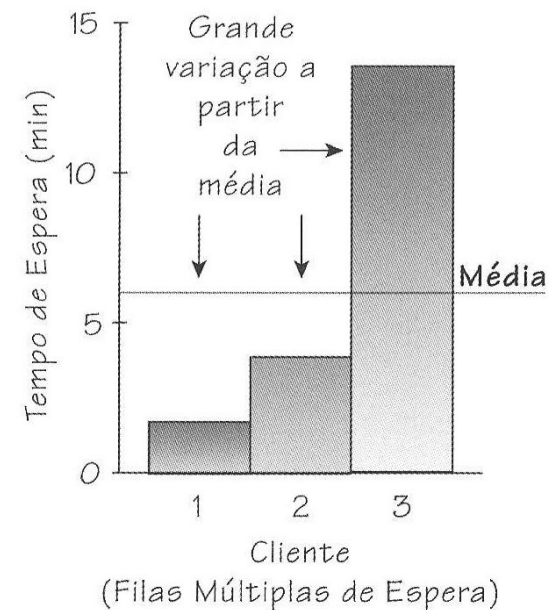
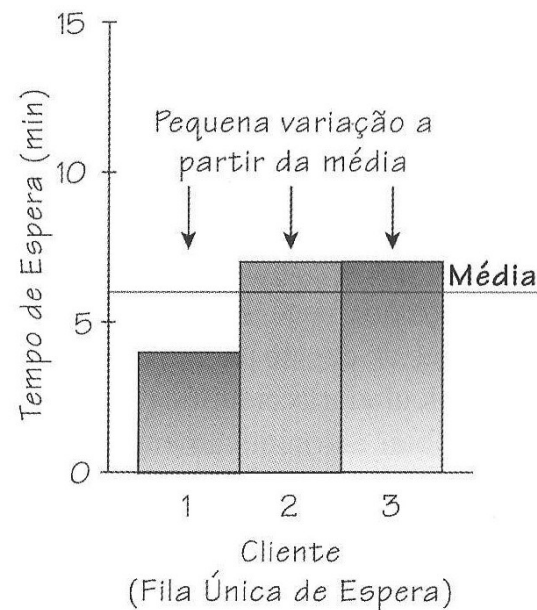
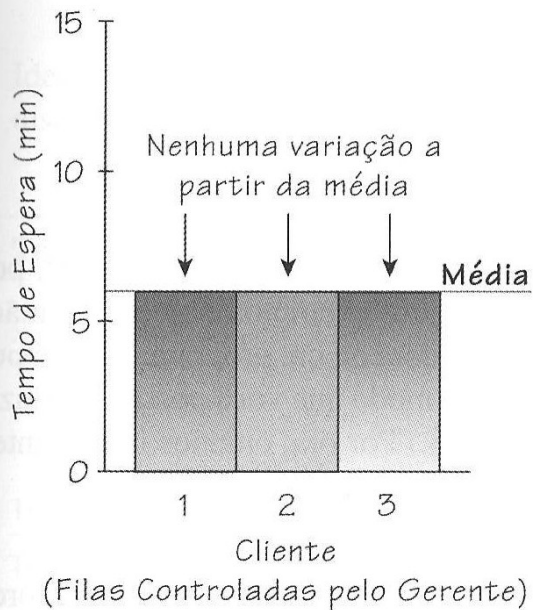


FIGURA 3-3 Tempos de Espera (min) de Clientes de Bancos

AMPLITUDE

- A amplitude de um conjunto de dados é a diferença entre o maior valor e o menor valor:

$$\text{amplitude} = (\text{valor máximo}) - (\text{valor mínimo})$$

- Essa é uma medida fácil de ser calculada.
- Porém, ao usar apenas os valores máximo e mínimo, não é tão útil quanto as outras medidas de variação que usam todos valores.

DESVIO PADRÃO AMOSTRAL

- O desvio padrão de um conjunto de valores amostrais é uma medida de variação dos valores em torno da média.
- Indica o desvio médio dos valores em relação à média.
- Fórmula do desvio padrão amostral:

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

- Fórmula que simplifica cálculos aritméticos:

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

PROPRIEDADES DO DESVIO PADRÃO

- O desvio padrão é uma medida da variação de todos valores a partir da média.
- O valor do desvio padrão (s):
 - É usualmente positivo.
 - Igual a zero quando todos valores dos dados são iguais.
 - Nunca é negativo.
- Maiores valores de s indicam maior variação.
- Valor de s pode crescer muito com a inclusão de um ou mais “outliers”.
- As unidades de s são as mesmas unidades dos dados originais.

CALCULANDO O DESVIO PADRÃO

- Calcule a média (\bar{x}).
- Subtraia a média de cada valor individual para obter uma lista de desvios ($x - \bar{x}$).
- Eleve ao quadrado cada uma das diferenças obtidas no passo anterior ($(x - \bar{x})^2$).
- Some todos quadrados obtidos no passo acima $\sum (x - \bar{x})^2$.
- Divida o total do passo anterior pelo total de valores presentes menos uma unidade ($n - 1$).
- Calcule a raiz quadrada do passo anterior.

DESVIO PADRÃO POPULACIONAL

- O desvio padrão da população (σ) utiliza o tamanho da população (N) no denominador:

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

VARIÂNCIA

- **Variância** de um conjunto de valores é uma medida da variação (dispersão) igual ao quadrado do desvio padrão.
- A **variância amostral** (s^2) é o quadrado do desvio padrão amostral (s).
- A **variância populacional** (σ^2) é o quadrado do desvio padrão populacional (σ).
- A variância amostral é considerada um **estimador não-viesado** da variância populacional:
 - Ao realizar várias vezes amostras aleatórias de uma população, os diferentes valores de s^2 tendem a se concentrar em torno do valor de σ^2 (sem superestimação ou subestimação).
- Unidades da variância são diferentes das unidades originais.

NOTAÇÃO E REGRA DE ARREDONDAMENTO

- s = desvio padrão amostral
- s^2 = variância amostral

- σ = desvio padrão populacional
- σ^2 = variância populacional

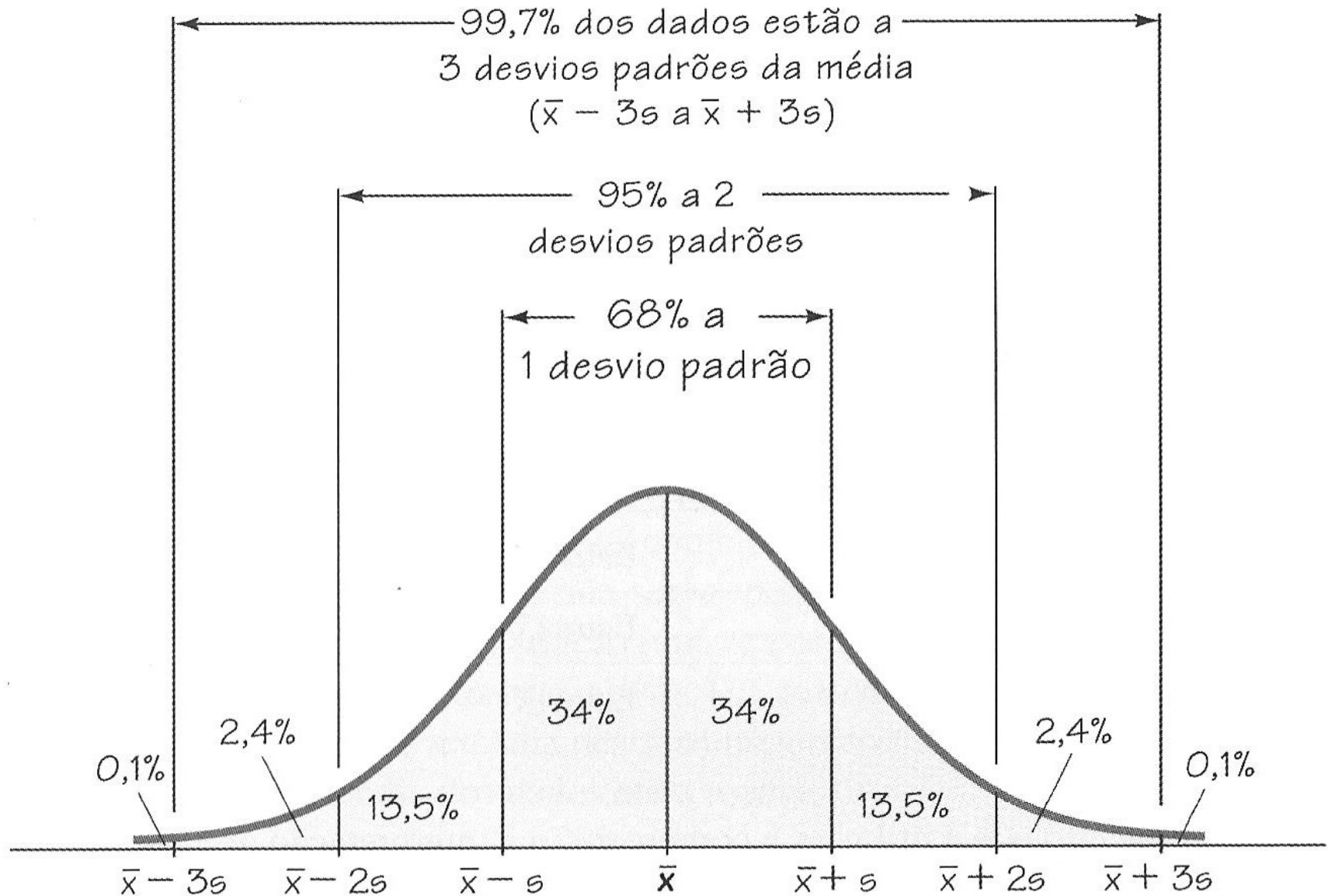
- SD = DP = desvio padrão (standard deviation)
- VAR = variância

- Como regra de arredondamento, use uma casa decimal a mais do que é apresentado no conjunto original de dados.

REGRA EMPÍRICA DA AMPLITUDE

- Desvio padrão mede a variação entre valores:
 - Valores muito próximos >>> desvios padrão pequenos.
 - Valores mais espalhados >>> desvios padrão maiores.
- A **regra empírica da amplitude** indica que para muitos conjuntos de dados, a grande maioria (95%) dos valores amostrais se localiza a 2 desvios padrões da média.
- Isso varia com tamanho amostral e natureza da distribuição.
- Desvio padrão (“grosseiro”) de dados amostrais:
$$s \approx \text{amplitude} / 4 \approx [(\text{valor máximo}) - (\text{valor mínimo})] / 4$$
- Valor amostral mínimo (usual) = média – (2 * desvio padrão)
- Valor amostral máximo (usual) = média + (2 * desvio padrão)

REGRA EMPÍRICA PARA DADOS COM FORMA APROXIMADA DE SINO (DISTRIBUIÇÃO NORMAL)



POR QUE DIVIDIR POR $n - 1$?

- Dividimos o desvio padrão amostral por $n - 1$, porque há apenas $n - 1$ valores independentes.
- Ou seja, dada uma média, apenas $n - 1$ valores podem ser associados a qualquer número, antes que o último valor seja determinado.
- Além disso, se s^2 fosse definido como a divisão por n , ele sistematicamente subestimaria o valor de σ^2 , o que é compensado pela diminuição do denominador.
- Vejam exercício 38 (pp. 88-89).

POR QUE EXTRAIR A RAIZ QUADRADA?

- Ao final do cálculo do desvio padrão, extraímos a raiz quadrada.
- Isso é realizado para compensar os quadrados que são estimados anteriormente.
- Ao calcular a raiz quadrada, o desvio padrão tem as mesmas unidades de medida dos dados originais.

COEFICIENTE DE VARIAÇÃO

- Por ter as mesmas unidades dos dados originais, o desvio padrão é mais fácil de entender do que a variância.
- Porém, com o desvio padrão, é difícil comparar a dispersão para valores de diferentes variáveis (ex.: peso e altura).
- **Coeficiente de variação** (CV) supera essa desvantagem, por não ter unidade específica, permitindo comparação das variações.
- O CV para um conjunto de dados amostrais ou populacionais não-negativos é expresso como um percentual e descreve o desvio padrão em relação à média:
 - Amostra: $CV = s/\bar{x} * 100\%$
 - População: $CV = \sigma/\mu * 100\%$

MEDIDAS DE POSIÇÃO RELATIVA

MEDIDAS DE POSIÇÃO RELATIVA

- As medidas de posição relativa permitem a comparação de valores de conjuntos de dados diferentes ou de valores dentro de um mesmo conjunto de dados.
- Os **escores z** permitem a comparação de valores de diferentes conjuntos de dados.
- Os **quartis** e **percentis** permitem a comparação de valores dentro do mesmo conjunto de dados, assim como entre diferentes conjuntos de dados.

ESCORES z

- Um escore z é obtido pela conversão de um valor para uma escala padronizada.
- O escore padronizado é o número de desvios padrões a que se situa determinado valor de x , acima ou abaixo da média:

- Amostra: $Z = \frac{x - \bar{x}}{s}$

- População: $Z = \frac{x - \mu}{\sigma}$

ESCORES z E VALORES NÃO-USUAIS

- Valores não-usuais são aqueles com escores z menores do que $-2,00$ ou maiores do que $+2,00$.
- Valores comuns: $-2 \leq \text{escore } z \leq 2$
- Valores não-usuais: $\text{escore } z < -2$ ou $\text{escore } z > 2$
- Sempre que um valor é menor do que a média, seu escore z correspondente é negativo.
- Escores z são medidas de posição, já que descrevem a localização de um valor (em termos de desvios padrões) em relação à média:
 - $z = 2$: valor está 2 desvios padrões acima da média.
 - $z = -3$: valor está 3 desvios padrões abaixo da média.

QUARTIS

- A **mediana** divide os dados ordenados em 2 partes iguais:
 - 50% dos valores de um conjunto de dados são iguais ou menores do que a mediana, e 50% são iguais ou maiores.
- Os **quartis (Q_1 , Q_2 e Q_3)** dividem os valores ordenados em 4 partes iguais:
 - Q_1 (primeiro quartil): separa os 25% inferiores dos 75% superiores.
 - Q_2 (segundo quartil): mesmo que a mediana; separa os 50% inferiores dos 50% superiores.
 - Q_3 (terceiro quartil): separa os 75% inferiores dos 25% superiores.

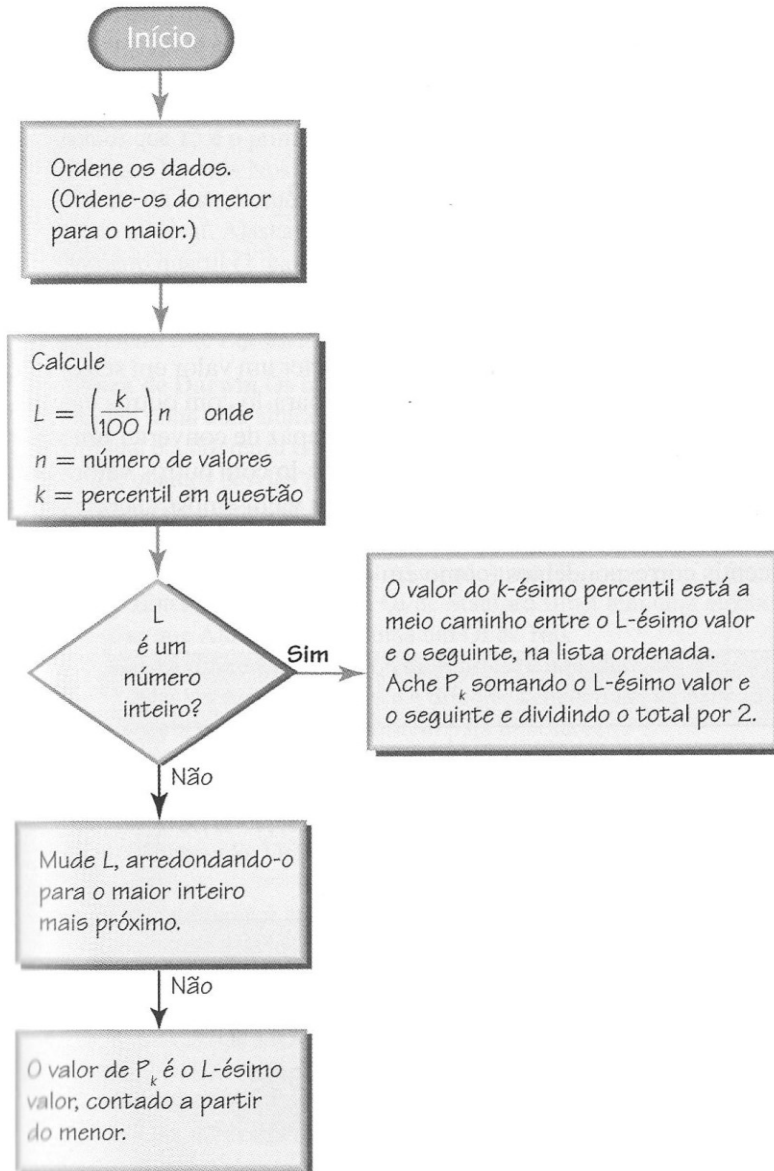
PERCENTIS

- Há 99 **percentis** (P_1, P_2, \dots, P_{99}) que dividem os dados ordenados em 100 grupos com cerca de 1% dos valores em cada um.
- Os quartis e percentis são exemplos de quantis, os quais dividem os dados em grupos com aproximadamente o mesmo número de valores.
- Utilize a seguinte fórmula, arredondando o resultado para o número inteiro mais próximo:

$$\text{percentil de } x = \frac{n^{\circ} \text{ valores menores do que } x}{n^{\circ} \text{ total de valores}} * 100$$

- Note que: $Q_1 = P_{25}$; $Q_2 = P_{50}$; $Q_3 = P_{75}$

CONVERTENDO PERCENTIS EM VALOR DE DADOS



– Sendo:

- n : número total de valores no conjunto de dados.
- k : percentil em uso (ex.: para o 25º percentil, $k=25$).
- L : localizador que dá a posição de um valor (ex.: para o 12º valor na lista ordenada, $L=12$).
- P_k : k -ésimo percentil (ex.: P_{25} é o 25º percentil).

FIGURA 3-6 Conversão do k -ésimo Percentil no Valor de Dado Correspondente

ESTATÍSTICAS DEFINIDAS POR QUARTIS E PERCENTIS

- Intervalo interquartil (IIQ) = $Q_3 - Q_1$
- Intervalo semi-interquartil = $(Q_3 - Q_1) / 2$
- Ponto médio dos quartis = $(Q_3 + Q_1) / 2$
- Intervalo percentílico 10–90 = $P_{90} - P_{10}$

ANÁLISE EXPLORATÓRIA DE DADOS (AED)

ANÁLISE EXPLORATÓRIA DE DADOS (AED)

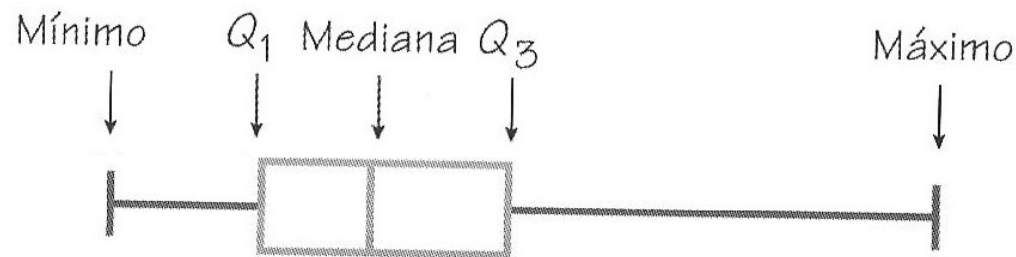
- Análise exploratória de dados é o processo de uso das ferramentas estatísticas (gráficos, medidas de centro, medidas de variação...) para investigação de conjuntos de dados com objetivo de se compreenderem suas características importantes.
- Podemos explorar características dos dados: centro (média, mediana); variação (desvio padrão, amplitude), distribuição (histogramas); *outliers*; mudança no tempo.
- Aqui serão discutidos os valores discrepantes (*outliers*) e o diagrama de caixa (*boxplot*).

VALORES DISCREPANTES (*OUTLIERS*)

- Valor *outlier* (valor extremo) é aquele que se localiza muito afastado de quase todos os demais valores.
- Estes valores podem ter efeito dramático sobre:
 - A média.
 - O desvio padrão.
 - A escala do histograma, de modo que a verdadeira natureza da distribuição pode ser totalmente obscurecida.
- *Outliers* podem ser erros: devem ser corrigidos ou ignorados
- *Outliers* podem ser corretos: devemos estudar seus efeitos, construindo gráficos e calculando estatísticas, com e sem *outliers*, buscando revelar importantes informações.

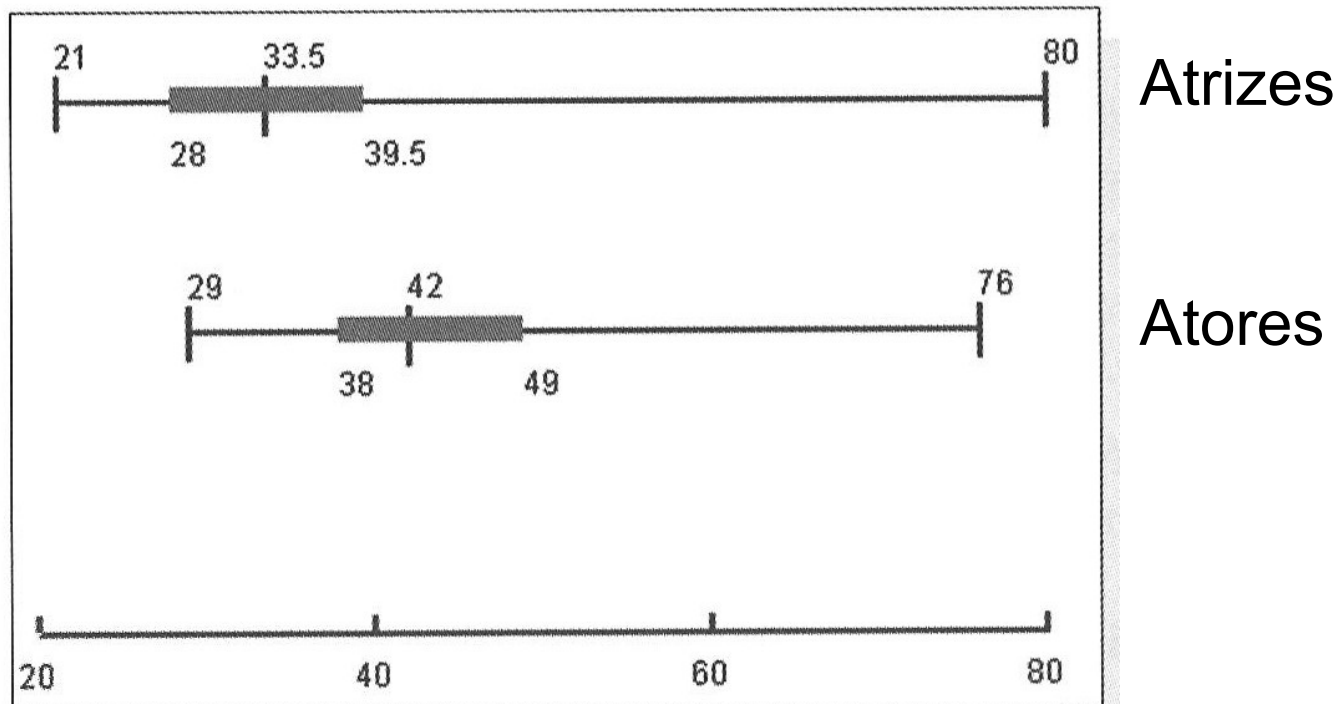
DIAGRAMAS DE CAIXA (*BOXPLOTS*)

- Para um conjunto de dados, o **resumo dos cinco números** consiste no valor mínimo, primeiro quartil (Q_1), mediana (Q_2), terceiro quartil (Q_3) e no valor máximo.
- **Diagrama de caixa** (diagrama de caixa e bigode) é um gráfico de um conjunto de dados que consiste em: (1) uma linha que se estende do valor mínimo ao valor máximo; (2) uma caixa com linhas traçadas no primeiro quartil (Q_1), na mediana (Q_2) e no terceiro quartil (Q_3).
- Os diagramas de caixa são úteis para revelar centro, dispersão, distribuição e *outliers*.



UTILIDADE DOS DIAGRAMAS DE CAIXA

- Diagramas de caixa não apresentam informação tão detalhada como histogramas e digramas de ramo e folhas.
- Porém, são úteis na comparação de dois ou mais conjuntos de dados, quando desenhados na mesma escala.
- *Boxplots* para idades dos melhores atores e atrizes:



DIAGRAMAS DE CAIXA MODIFICADOS

- Diagramas de caixa modificados representam *outliers* com símbolos especiais (asteriscos).
- Lembrando que $IQ = Q_3 - Q_1$, um valor é *outlier* se está:
 - Acima de Q_3 por uma quantidade maior do que $1,5 \times IQ$.
 - ou
 - Abaixo de Q_1 por uma quantidade maior do que $1,5 \times IQ$.
- A linha sólida horizontal se estende apenas até o menor valor dos dados que não são *outliers* e até o maior valor dos dados que não são *outliers*.

