ESCOLA DE ENGENHARIA MAUÁ

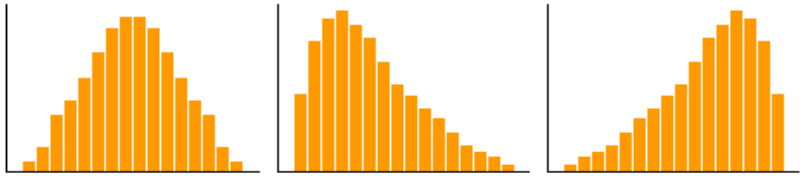
EFB803 Estatística

2020
Aula 03

(Análise descritiva: assimetria, histograma e boxplot)

Forma da distribuição

- Conhecendo as **medidas de posição** (ou tendência central) e de **dispersão** (ou variabilidade) de uma distribuição, interessa agora avaliar **a forma aproximada da distribuição dos dados**.

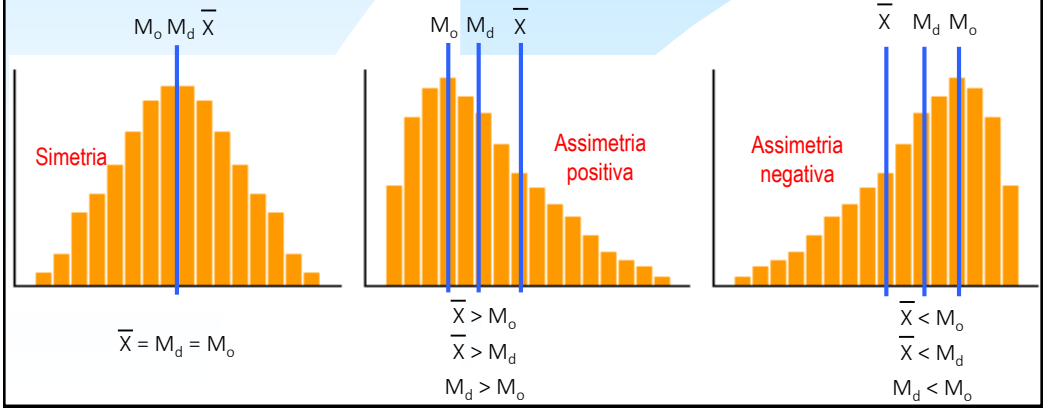


Há duas maneiras de se avaliar a forma da distribuição dos dados:

i) quantificar um coeficiente de assimetria e/ou,
ii) análise visual (histograma).

Forma da distribuição (análise visual)

- Entender a forma como os valores de uma variável se distribuem é bastante comum em controle de qualidade:
  - ✓ Distribuição simétrica: processo em controle
  - ✓ Distribuição assimétrica: em geral, indicam causas sistemáticas de falhas (ajustes do equipamento, por exemplo)



Forma da distribuição

- Existem várias "fórmulas" na literatura para quantificar a assimetria de uma variável;
- O **coeficiente de assimetria** é um valor abstrato e adimensional, que permite comparar a forma geral de várias distribuições entre si;
- Quando a distribuição é **simétrica**, o coeficiente de assimetria é **nulo**. Quando **não é nulo**, diz-se que a distribuição é **assimétrica**;
- O sinal (+ ou -) indica o ramo da distribuição (direita ou esquerda) mais assimétrico.

Forma da distribuição: **Coeficiente de assimetria (A)**

Coeficiente A de assimetria **de Pearson**

$$A = \frac{\bar{x} - M_o}{s}$$

Quando **não** se tem a moda (ou quando há mais de uma moda) e a distribuição da variável parecer ser razoavelmente simétrica, essa expressão fornece uma boa aproximação

$$A \approx \frac{3(\bar{x} - M_d)}{s}$$

são as "fórmulas" mais simples que existem para avaliar a assimetria de uma variável quantitativa. Existem coeficientes melhores

Critério de classificação pelo Coeficiente de Pearson

$|A| < 0{,}15$ (simetria)

$0{,}15 \le |A| \le 1$ (assimetria leve a moderada)

$|A| > 1$ (assimetria forte)

Não calculam as expressões mostradas acima

Forma da distribuição: **Coeficiente de assimetria (A_3)**

- Nos softwares de uso profissional outro critério é utilizado para calcular o coeficiente de assimetria (*skewness*, em inglês):

$$A_3 = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

Curiosidade: o índice 3 subscrito se refere ao uso do momento de 3ª ordem no cálculo

=distorção(Axx:Axx)

Critério de classificação pelo Coeficiente A_3

- ✓ **Valores muito próximos de 0 indicam simetria;**
- ✓ **Valores afastados de 0, mas entre -1 e +1 indicam leve assimetria**
- ✓ **Fora desse intervalo a assimetria é de moderada a forte intensidade**

Exercício (Minitab)

A PStores é uma das divisões de uma cadeia de lojas de roupas femininas. Foi realizada uma promoção recentemente na qual cupons de desconto foram enviados aos clientes das demais divisões e alguns clientes responderam uma pesquisa (quantidade de itens comprados, valor da compra, tipo de cartão utilizado, gênero, estado civil e idade) após realizarem suas compras na PStores.

Como os cupons promocionais não foram enviados aos clientes regulares da rede, a administração considera as negociações feitas para pessoas apresentando cupons promocionais como vendas que de outro modo, não teriam sido efetuadas. Uma amostra de 100 clientes foi registrada e os dados estão disponíveis na aba "***Dados***" do arquivo **Aula03.xlsx**. Uma legenda das variáveis é apresentada na outra aba da planilha.

Avalie a forma da distribuição das variáveis "Valor gasto" e "Idade", calculando o coeficiente A3.

Valor da compra: $A_3 \approx 1{,}71$ (forte assimetria)

Idade: $A_3 \approx 0{,}52$ (leve assimetria)

Forma da distribuição: Histograma

- Com os dados resumidos em uma distribuição de frequências (aula passada) tem-se um resumo mais compacto deles;
- A versão mais simples do histograma é, a partir de uma tabela de distribuição de frequências, construir um gráfico de colunas de modo que elas fiquem dispostas "coladas" umas nas outras;

Principal objetivo: avaliar se uma variável quantitativa apresenta simetria ou assimetria na distribuição de seus valores.

Exemplo

Distribuição de frequências do rendimento mensal de uma aplicação financeira.

Rendimento mensal (%)	Freqüência
0,60 ⊢ 0,70	4
0,70 ⊢ 0,80	2
0,80 ⊢ 0,90	4
0,90 ⊢ 1,00	2

4 classes

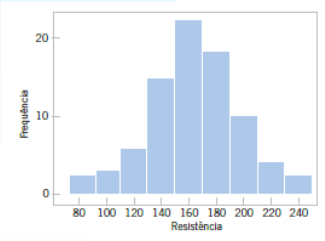
Cuidados na construção do histograma

Se a tabela de frequências tiver um número muito baixo ou muito alto de classes, o histograma não será informativo para avaliar da forma da distribuição da variável X de interesse. Então, ele deve ser construído usando:

Quantidade de classes do gráfico: $k \cong \sqrt{n}$

Ideal de 5 a 15 classes

n representa o tamanho da amostra e k é um número inteiro arredondado convenientemente



exemplo com k = 9 classes, sendo cada classe com amplitude de 20 unidades (convenientemente arredondado)

Exercício (Minitab)

Para a amostra de 100 clientes que compraram nas lojas Pstores, já foi calculado o coeficiente de assimetria de algumas variáveis. Considere X = Valor da compra e Y = Idade. Para eles, responda:

a) Um histograma com quantas classes seria razoável para analisar a forma de X e de Y?

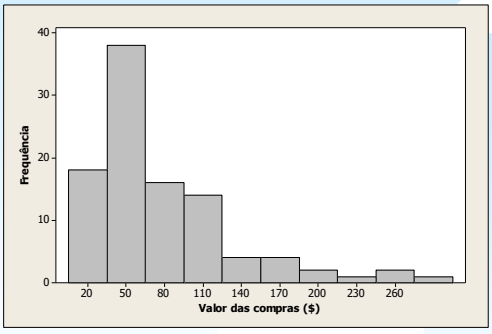
$k = \sqrt{100} = 10$

b) Construa o histograma de cada variável do exercício.

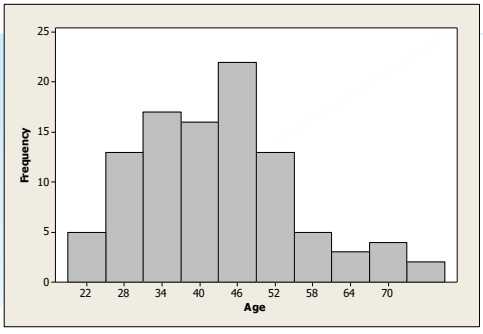
c) Analise os histogramas obtidos junto com os coeficientes A_3 calculados anteriormente.

(Pelo gráfico, a distribuição de X parece apresentar forte assimetria positiva (A_3 = 1,71) ; O histograma de Y sugere de leve a moderada assimetria positive (A_3 = 0,52)

X = Valor da compra

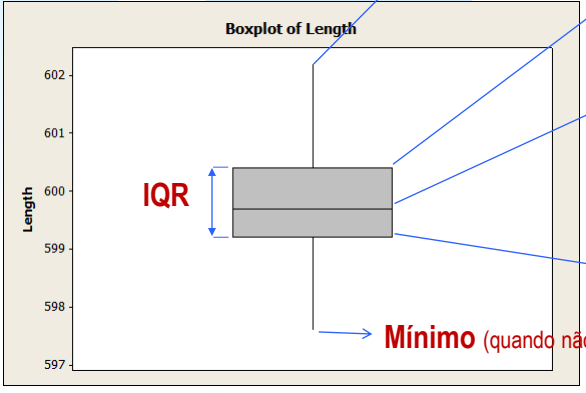


Y = Idade



Boxplot (diagrama de caixas)

- Literalmente, é um gráfico de caixa.
- Pode ser representado na vertical (a largura não tem significado) ou na horizontal (a altura não tem significado);
- Possui algumas utilidades:



Máximo (quando não há valores atípicos)

Q_3 (3º quartil)

Mediana (Q_2 ou 2º quartil)

IQR

Q_1 (1º quartil)

Mínimo (quando não há valores atípicos)

$IQR = Q_3 - Q_1$
(intervalo interquartil)

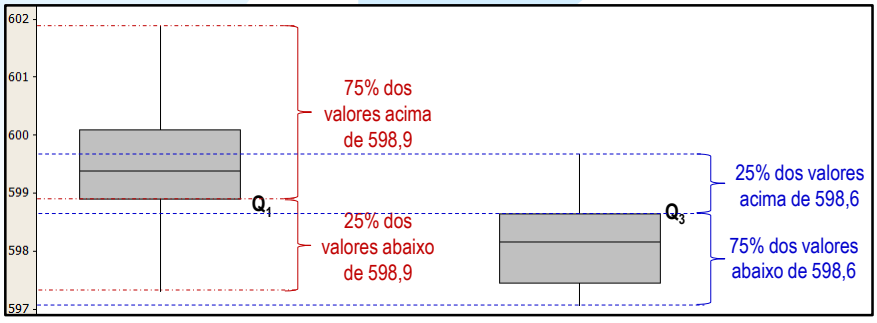
Conceitos:

Q_1: estatística que divide 25% da amostra com os menores valores dos 75% da amostra com os maiores valores;

Mediana: divide metade da amostra com as menores observações da metade com as maiores observações;

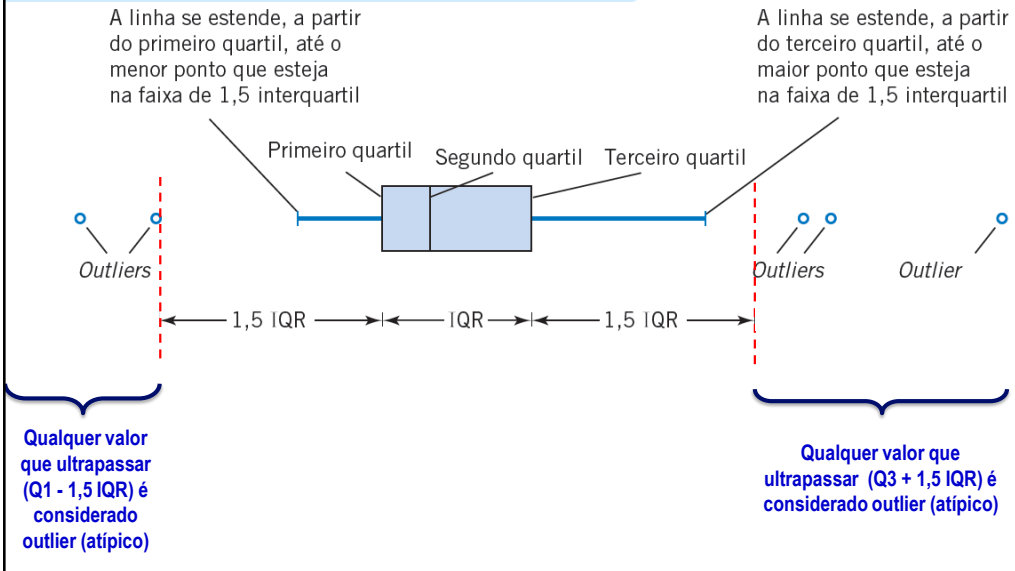
Q_3: estatística que divide 75% da amostra com os menores valores dos 25% da amostra com os maiores valores.

Exemplo. Volume mensal desviado (em milhares de reais) de uma estatal por dois partidos políticos nos últimos 120 meses. Os valores são hipotéticos, pois sabemos que partido nenhum desvia dinheiro público!!!



Partido Vermelho **Partido Azul**

Box-plot (construção)



Exercício (Minitab)

Para os dados dos clientes que compraram nas lojas PStores, faça o que se pede abaixo:

a) Compare a idade dos clientes em função do estado civil. Calcule as principais medidas resumo e interprete. Qual grupo parece ser mais velho?

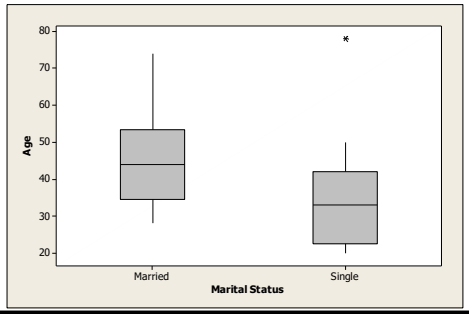
Descriptive Statistics: Age

Variable	Marital Status	N	N*	Mean	StDev	Q1	Median	Q3	Skewness
Age	Married	84	0	44.50	11.38	34.50	44.00	53.50	0.56
	Single	16	0	35.63	15.04	22.50	33.00	42.00	1.49

b) Faça a mesma comparação, porém visualmente agora.

Item a: Tanto em média como em mediana, os casados parecem ser mais velhos do que os solteiros.

Item b: Pelo box-plot, é mais fácil ver que metade dos casados tem mais idade do que 75% dos solteiros, indicando, no geral, que eles parecem ser mais velhos.



Exercício (Minitab)

Para os dados dos clientes que compraram nas lojas PStores, faça o que se pede abaixo:

a) Compare o valor gasto nas compras entre casados e solteiros. Calcule as principais medidas resumo e interprete. Há indícios que algum grupo gaste mais?

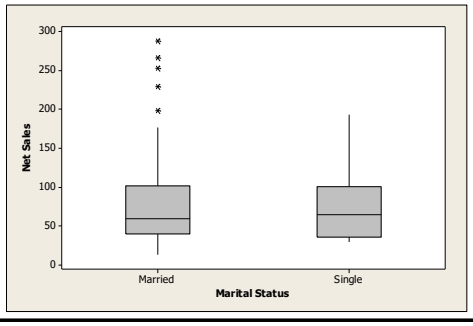
Descriptive Statistics: Net Sales

Variable	Marital Status	N	N*	Mean	StDev	Q1	Median	Q3	Skewness
Net Sales	Married	84	0	78.03	57.67	39.60	59.00	101.90	1.73
	Single	16	0	75.4	45.2	36.1	64.5	100.4	1.35

b) Faça a mesma comparação, porém visualmente agora.

Item a: Tanto em média como em mediana, o valor gasto nas compras parecem ser próximos entre si. Além disso a variabilidade é alta em ambos os grupos.

Item b: Pelo box-plot, é mais fácil ver o comportamento bem parecido entre os dois grupos. Além disso, nota-se alguns valores atípicos no grupo dos casados, fazendo com que a média fique um pouco mais elevada e alterando também o desvio padrão desse grupo.

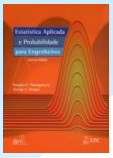


Estudo recomendado

- Fazer o(s) exercício(s) que não foram finalizados na aula

Cap. 1

Seção 1.3 a 1.6 e seus respectivos exercícios



Cap. 6

Seções 6.3 e 6.4 e seus respectivos exercícios

- Estudar o material complementar (pdf no Moodle) sobre como calcular os quartis e esboçar o boxplot manualmente.