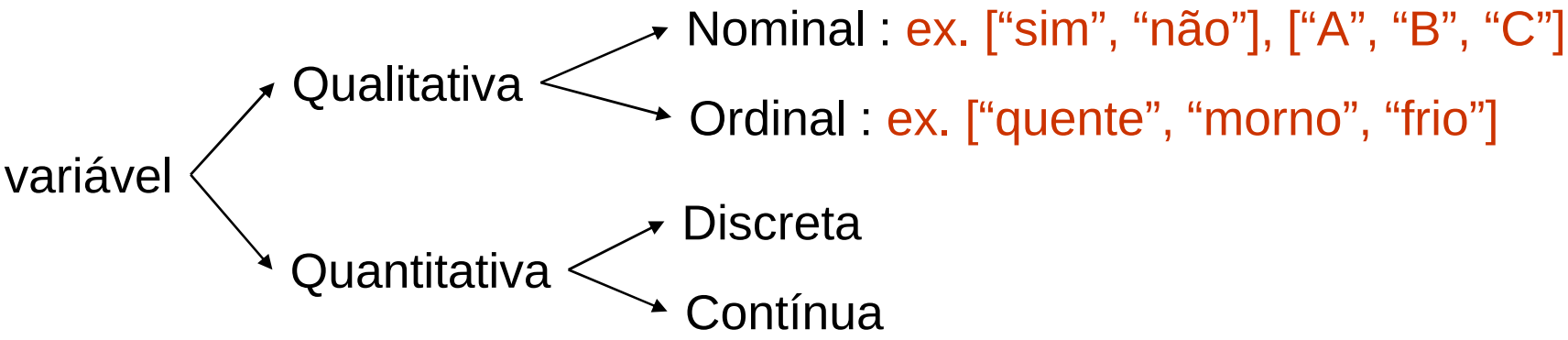
Bioestatística F

Estatística Descritiva

Enrico Antonio Colosimo
Depto. Estatística – UFMG
http://www.est.ufmg.br/~enricoc/

Descrição de Dados

- Variável: característica de interesse



Variáveis : resposta e covariáveis

Definições

- ***Variáveis quantitativas discretas*** – podem ser vistas como resultantes de contagens, assumindo assim, em geral, valores inteiros.

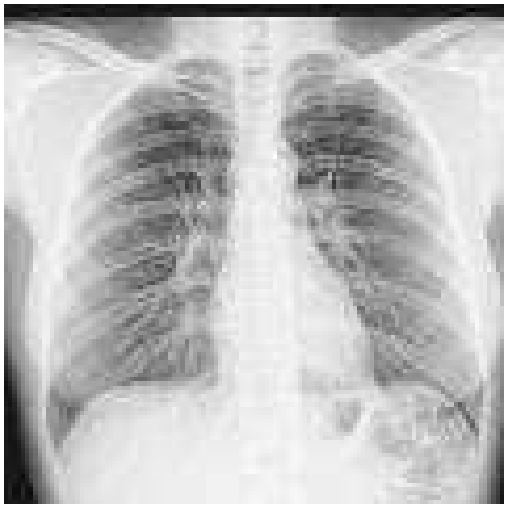
- ***Variáveis quantitativas contínuas*** – assumem valores em intervalos dos números reais e, geralmente, são provenientes de uma mensuração.

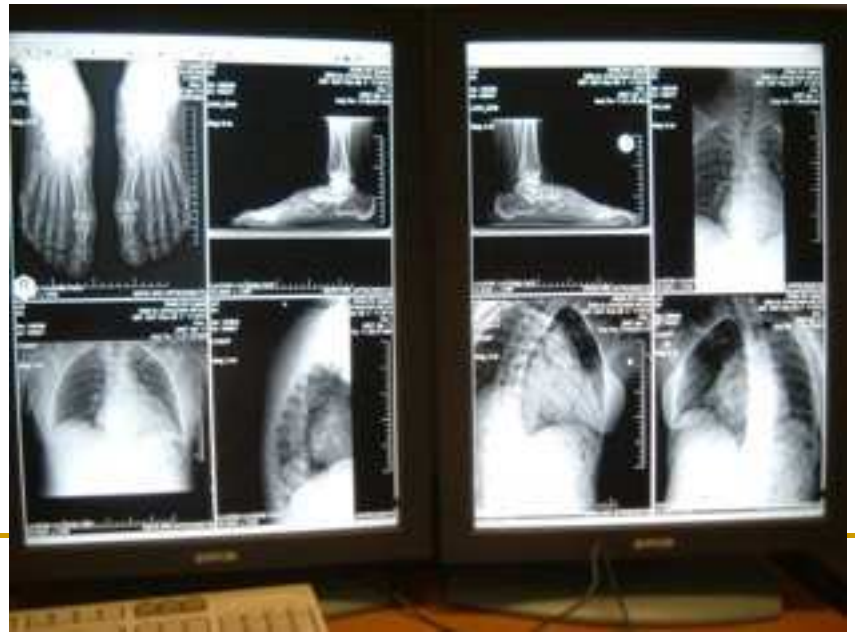
Estatística Descritiva

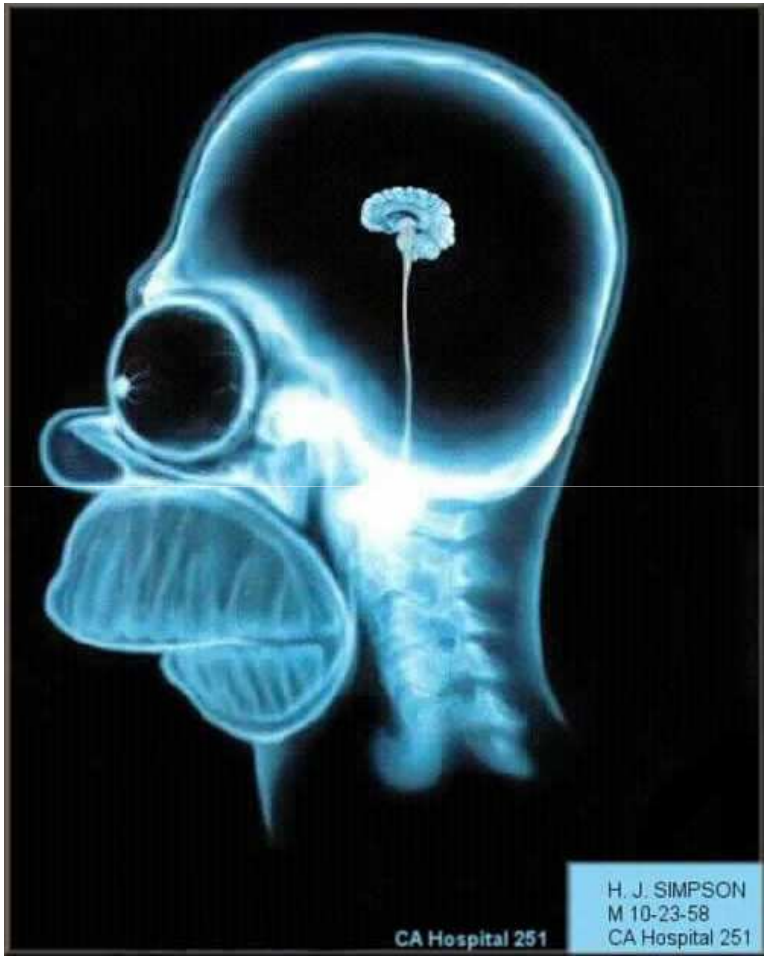
- Descobrindo o óbvio?

O que parece ser simplesmente uma foto de grãos de café, porém não é. Existe a face de um homem entre os grãos. Segundo dizem, se você conseguir encontrar o homem em 3 segundos ou menos a sua parte direita do cérebro é mais desenvolvida do que a maioria.

Interpretando Gráficos







Descrição e Apresentação de Dados (Estatística Descritiva)

- **Exemplo:** Em 1969 foi realizado um estudo na população de Honolulu. Para 7.683 indivíduos foram pesquisadas as seguintes variáveis: nível educacional, peso (*kg*), altura (*cm*), idade (anos), glicemia (*mg/dL*), colesterol sérico (*mg/dL*) e pressão sistólica (*mmHg*). Cada indivíduo foi classificado quanto ao hábito de fazer exercício físico e uso de tabaco.

Base de Dados

ID	EDUCATIONAL LEVEL	WEIGHT (KG)	HEIGHT (CM)	AGE	SMOKING STATUS	PHISICAL ACTIVITY AT HOME	BLOOD GLUCOSE	SERUM CHOLESTEROL	SYSTOLIC BLOOD PRESSURE
1	2	70	165	61	1	1	107	199	102
2	1	60	162	52	0	2	145	267	138
3	1	62	150	52	1	1	237	272	190
4	2	66	165	51	1	1	91	166	122
5	2	70	162	51	0	1	185	239	128
6	4	59	165	53	0	2	106	189	112
7	1	47	160	61	0	1	177	238	128
8	3	66	170	48	1	1	120	223	116
9	5	56	155	54	0	2	116	279	134
10	2	62	167	48	0	1	105	190	104
11	4	68	165	49	1	2	109	240	116
12	1	65	166	48	0	1	186	209	152
13	1	56	157	55	0	2	257	210	134
14	2	80	161	49	0	1	218	171	132
15	3	66	160	50	0	2	164	255	130
16	4	91	170	52	0	2	158	232	118
17	3	71	170	48	1	1	117	147	136
18	5	66	152	59	0	2	130	268	108
19	1	73	159	59	0	2	132	231	108
20	4	59	161	52	0	1	138	199	128
21	1	64	162	52	1	1	131	255	118
22	3	55	161	52	1	1	88	199	134

Introdução aos Gráficos

- Gráficos
  - Disco/Torta/Pizza
  - Barras (variável versus frequência)
  - Histograma (polígono de frequência)
  - Box-plot

- Histograma (original): variável versus densidade. Possui área total igual a 1.

Retângulos contíguos com área igual à frequência relativa (densidade de frequência). As densidade de cada faixa podem ser obtidas dividindo-se a frequência relativa pela amplitude da faixa.

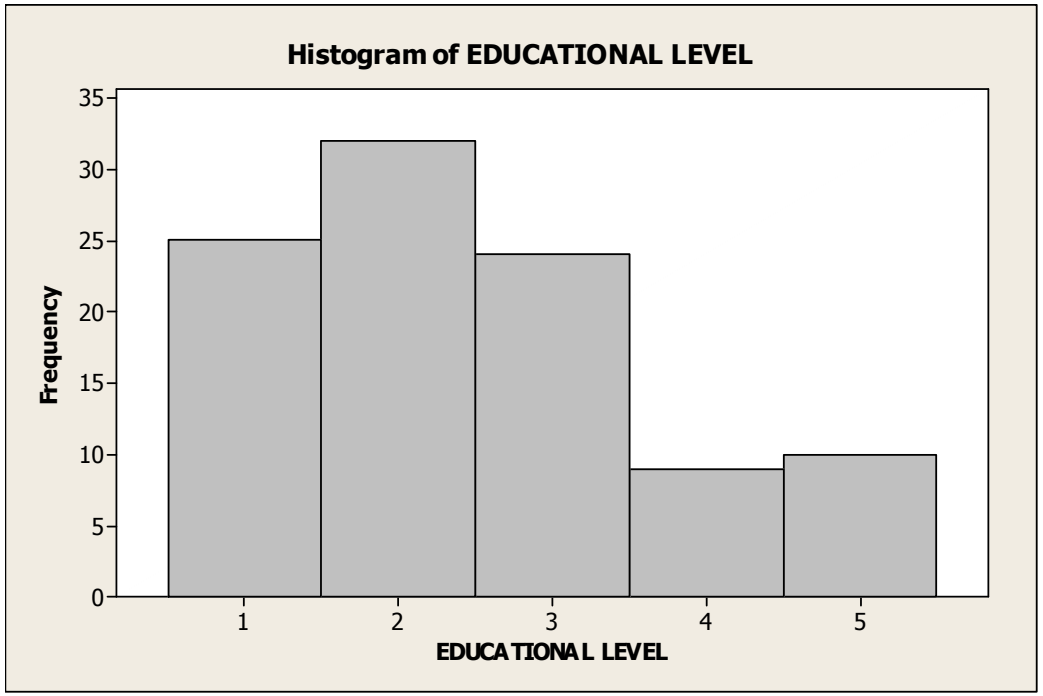
Tabela de Frequência

Variável	n_i	f_i	fac
valor 1			
valor 2			
⋮			
valor p			
Total	n=	1	

fac = frequências acumuladas

x	**n_i**	**f_i**	**fac**
A	38	0,38	0,38
B	20	0,20	0,58
C	42	0,42	1,00
Total	**100**		

Educational Level



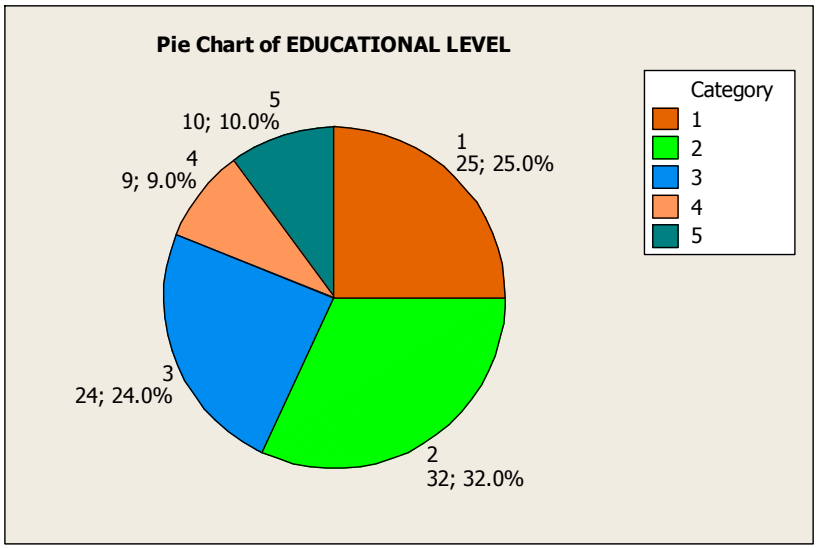
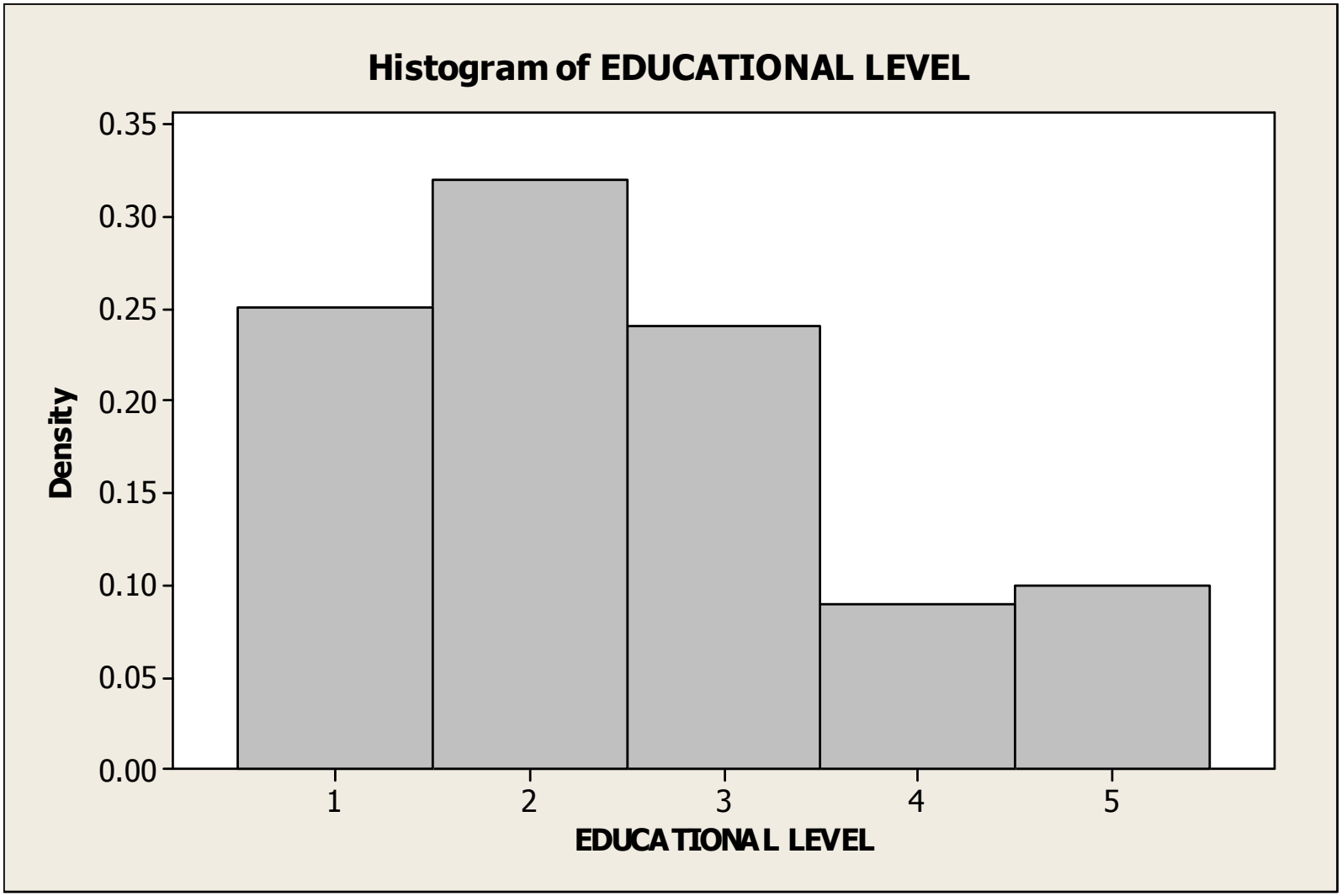
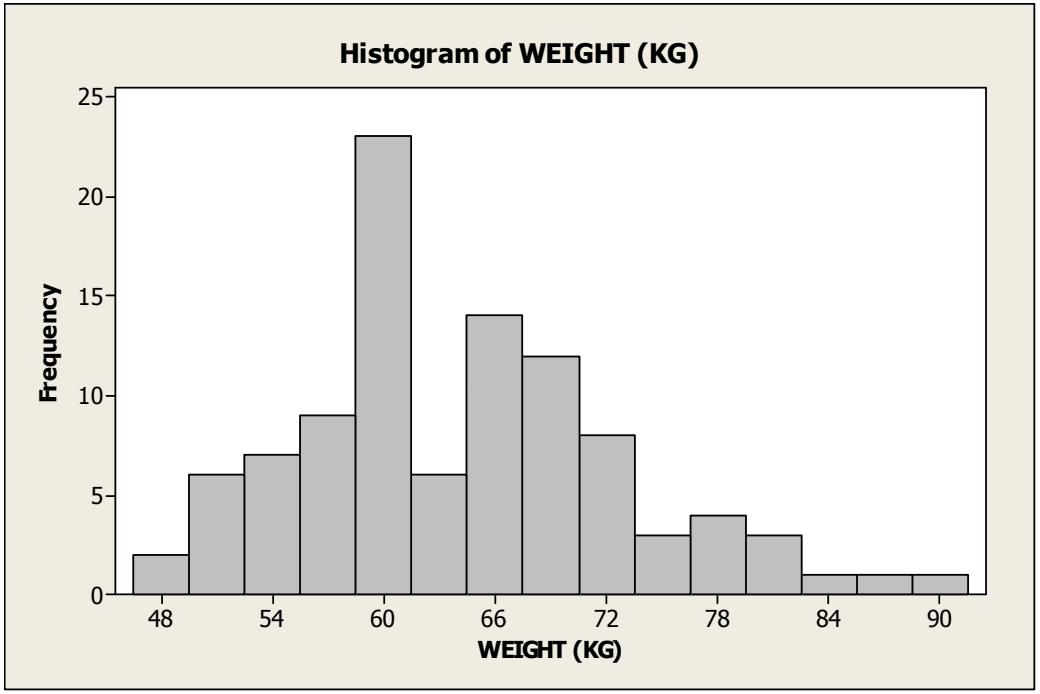


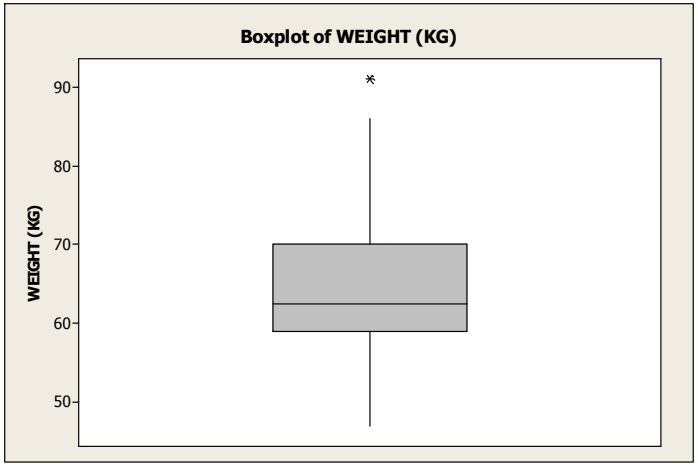
Gráfico de Densidade



Weight (Kg)



Weight (kg)	
Média	64.22
Mediana	62.5
Desvio padrão	8.6
Mínimo	47
Máximo	91
Contagem	100



Medidas de Posição (Tendência Central)

1. Média (amostral)
   Sejam x_1, x_2, ..., x_n observações da variável X

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

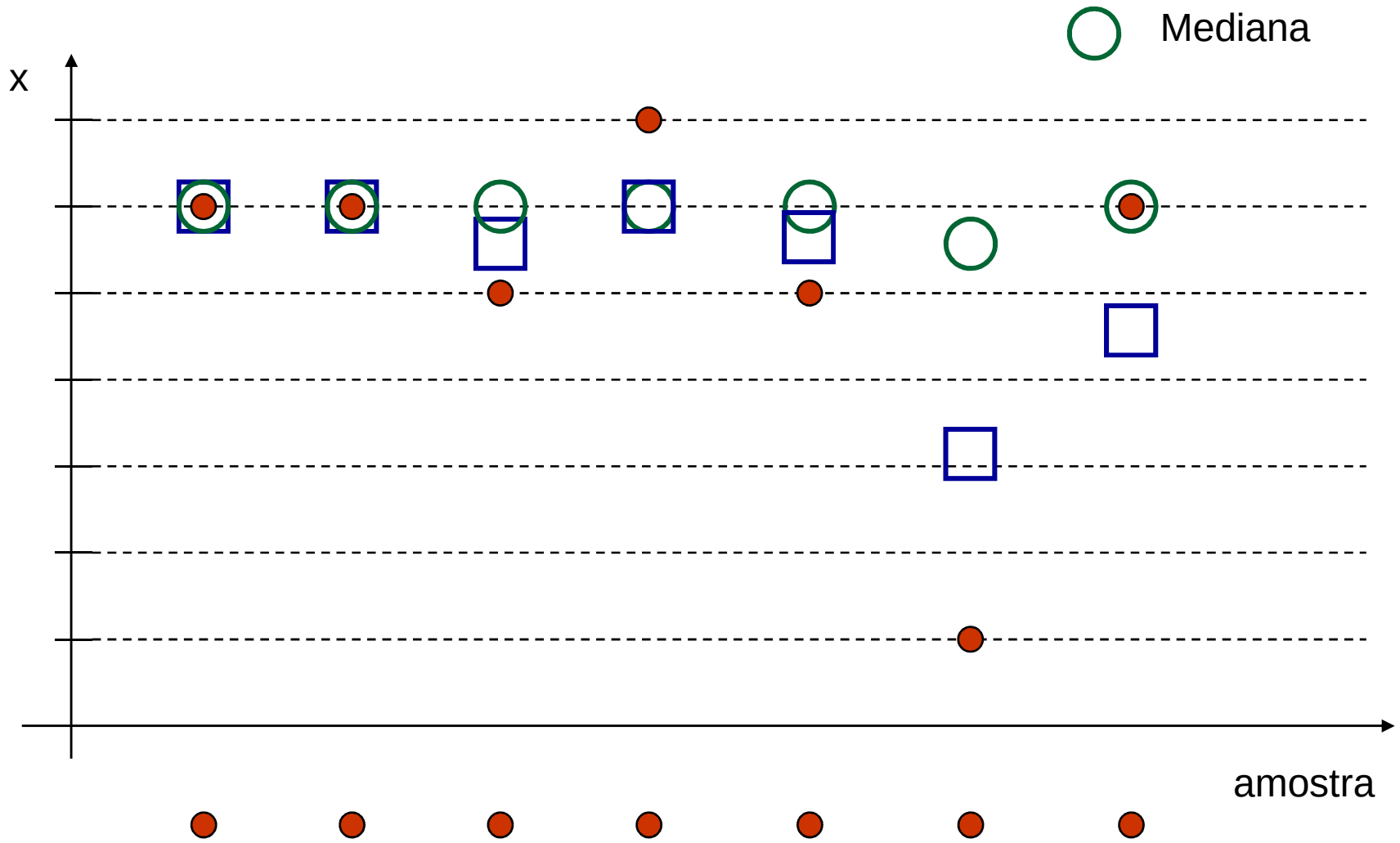
ou na forma:

$$\overline{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \sum_{i=1}^{k} \frac{n_i}{n} x_i$$

2. Mediana
   md: é o valor que ocupa a posição central dos dados ordenados

   também representada como: $\tilde{x}$

Média versus Mediana



Medidas de Dispersão

1. Amplitude: é a diferença entre o maior e o menor valor do conjunto de dados.

2. Variância (amostral)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

3. Desvio Padrão: S

Percentis e Quartis

- **Percentil**: o percentil de ordem 100p% de um conjunto de valores dispostos em ordem crescente é um valor tal que pelo menos (100p)% das observações são menores ou igual a ele **e**, pelo menos 100(1-p)% são maiores ou igual a ele.

- **Quartis**: os percentis de ordem 25, 50 e 75 são chamados quartis. Representam-se por Q_1, Q_2 (mediana) e Q_3.

OBS: Q_1 deixa pelo menos 25% dos dados abaixo e pelo menos 75% acima dele

Regras Práticas (n observações)

P_k, : k-ésimo percentil; valores ordenados da amostra

1. **Se nk/100 é inteiro: P_k é obtido tomando a média deste valor com o vizinho superior.**
2. **Se nk/100 não é inteiro: P_k é aquele imediatamente superior.**

Regra utilizada pela "maioria"dos softwares (SPSS e outros):

Tomar (n+1)k/100: se for inteiro é o próprio valor e se não for, fazer interpolação entre os vizinhos.

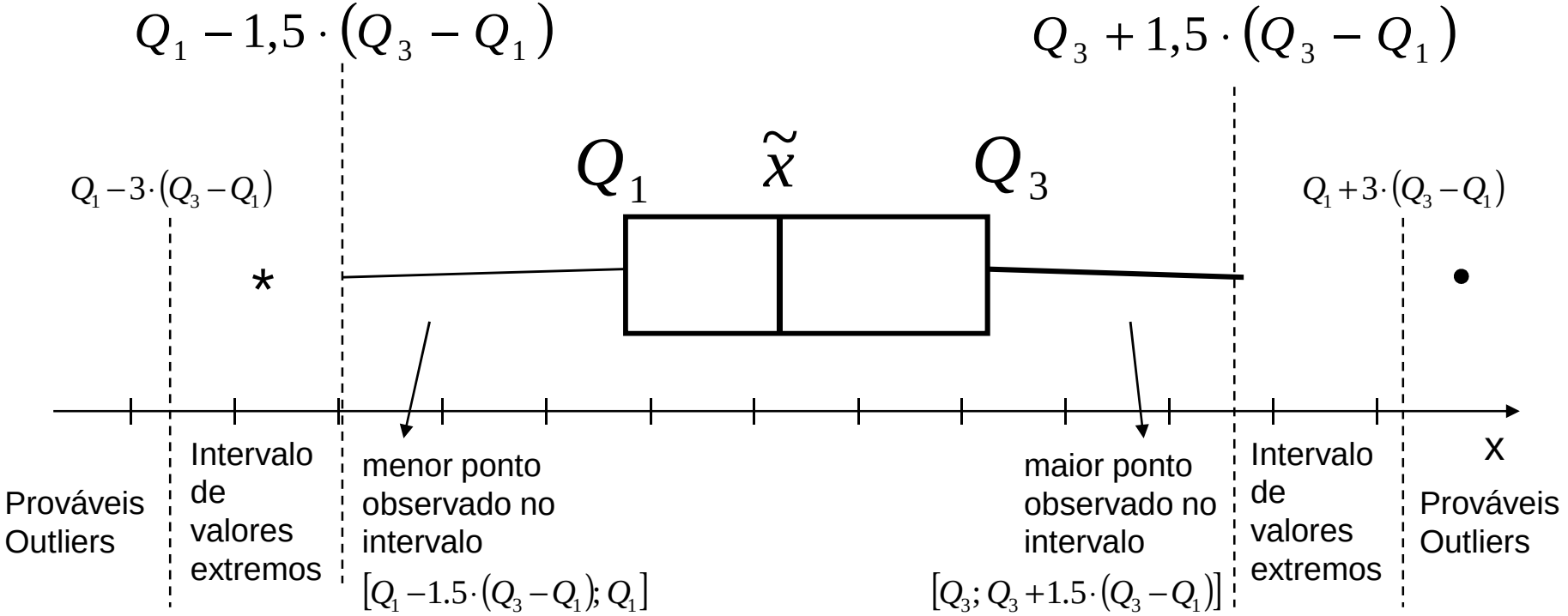
Exemplos: Percentis

Dados: 1,2,3,4,5,6,7,8,9,10.

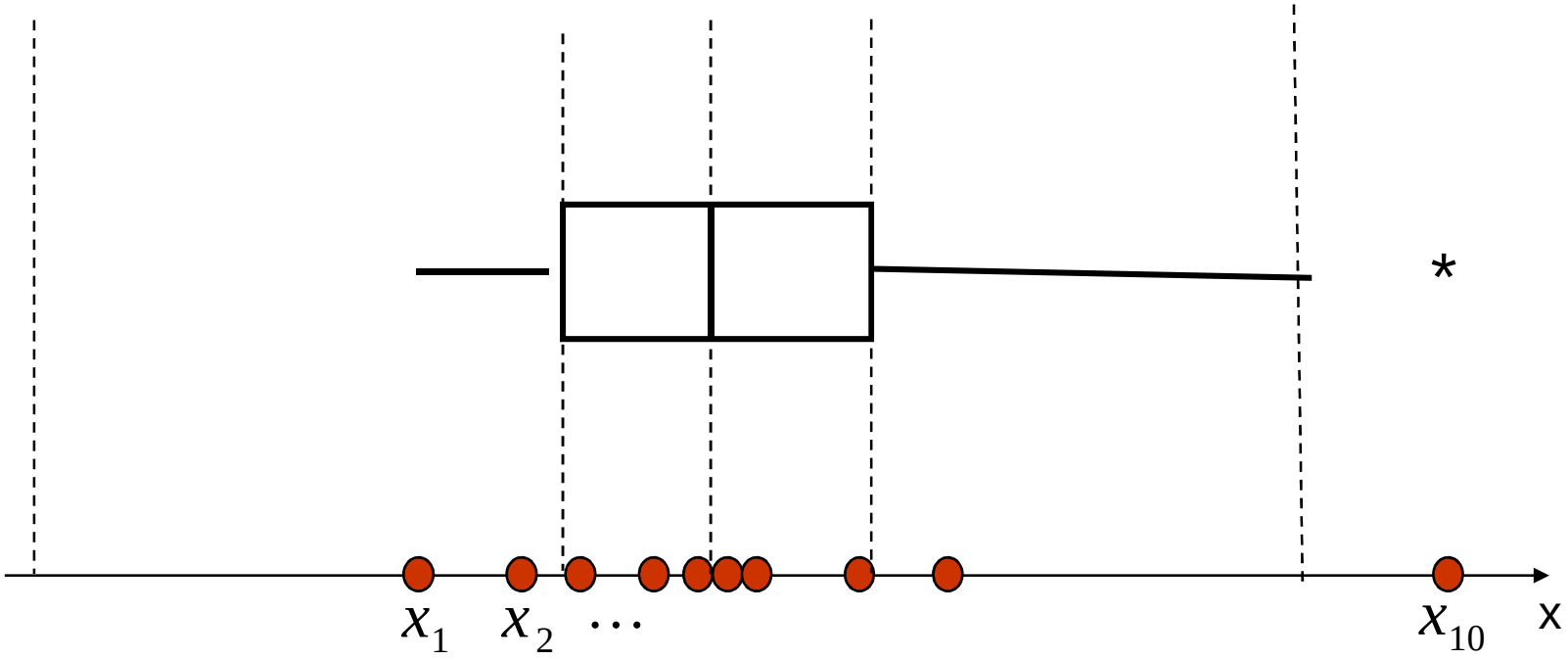
Encontrar: $\mathbf{P_{10}}$, $\mathbf{P_{50}}$, $\mathbf{Q_3} = \mathbf{P_{75}}$

BoxPlot

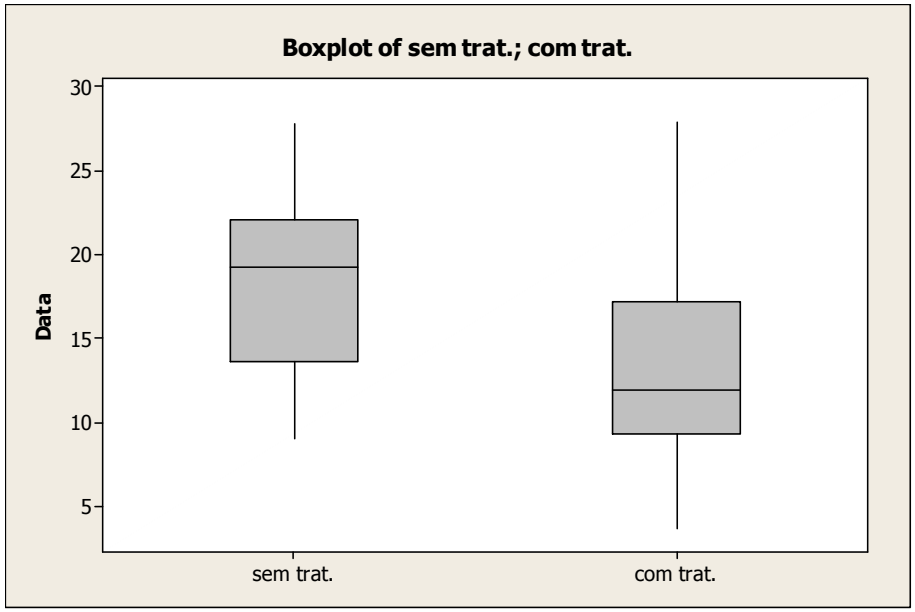
- É um gráfico que apresenta simultaneamente várias características de dados: locação, dispersão, simetria e presença de observações discrepantes (“outliers”)



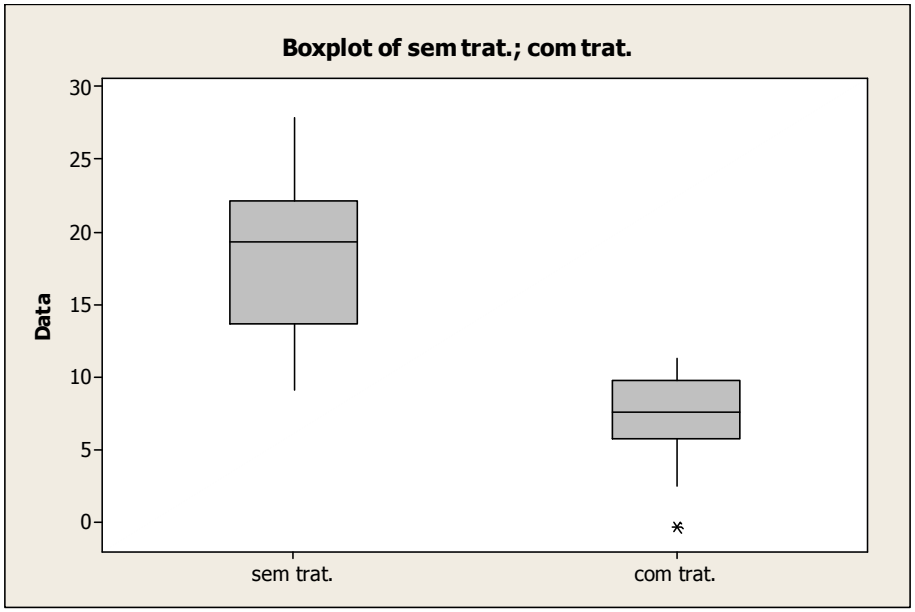
BoxPlot



Comparação de Grupos



Droga 1



Droga 2

Causas do aparecimento de *outliers*

- Leitura, anotação ou transcrição incorreta dos dados.
- Erro na execução do experimento ou na tomada da medida.
- Mudanças não controláveis nas condições experimentais ou dos pacientes.
- Características inerente à variável estudada (por exemplo, grande instabilidade do que está sendo medido)

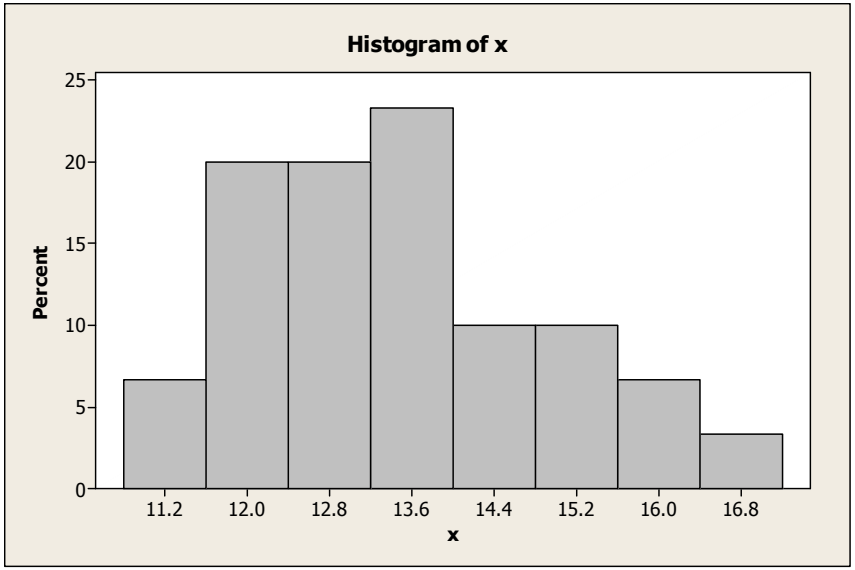
Exercícios

- Foram feitas medidas em operários da construção civil a respeito da taxa de hemoglobina no sangue (gramas/cm3)

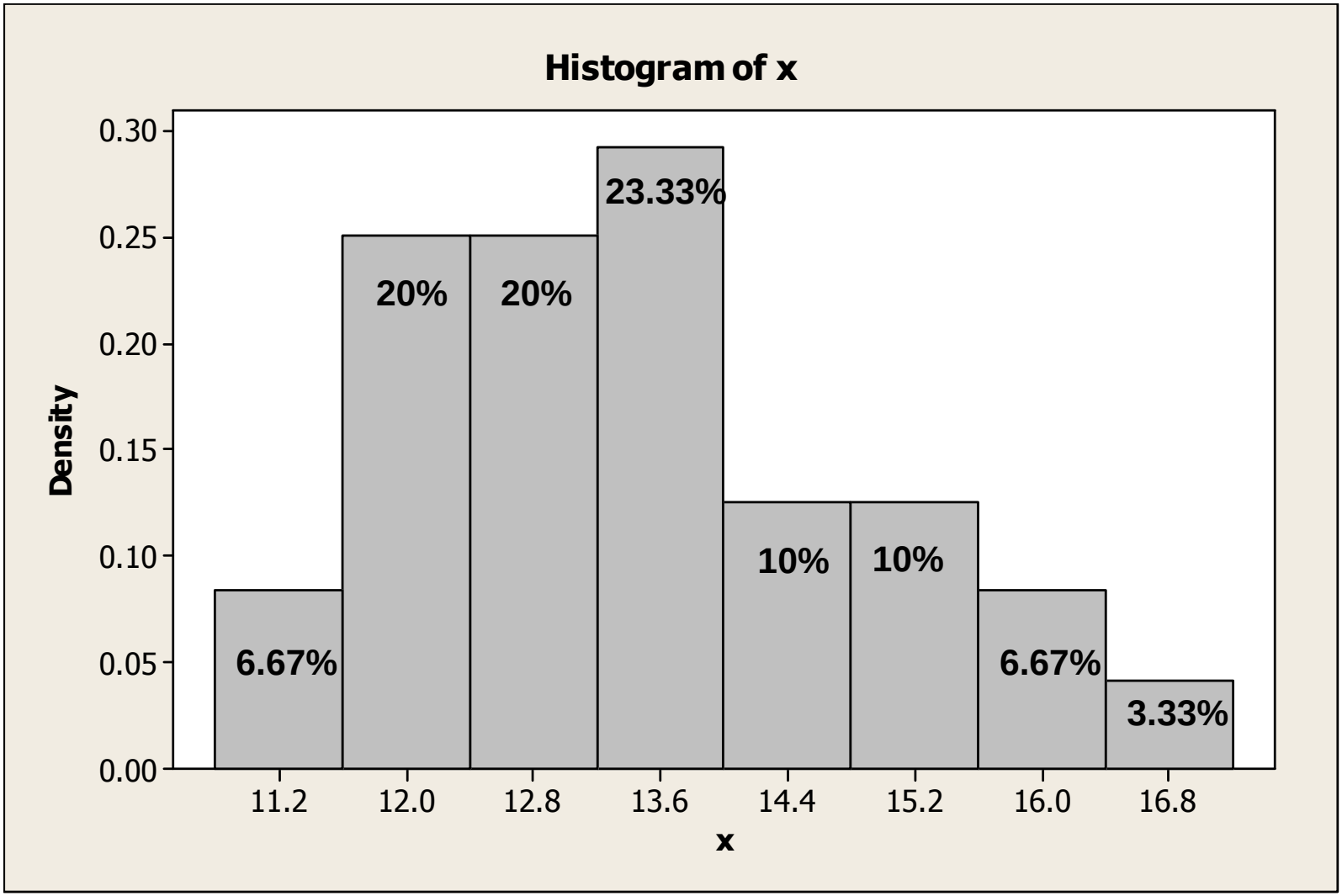
A partir do histograma gerado obtenha:

a) A mediana (Q_2)

b) Os quartis (Q_1 e Q_3)

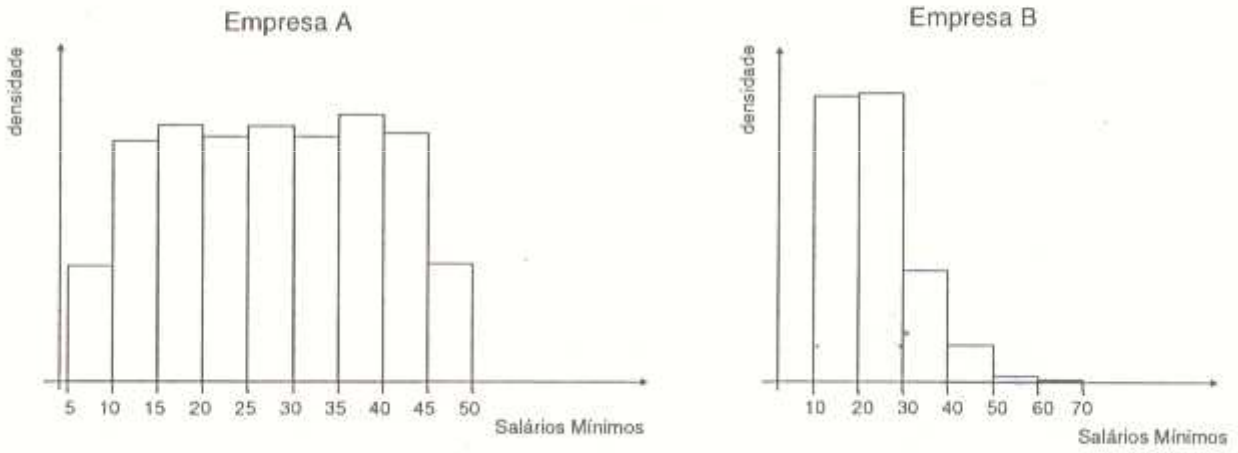


Histograma (Exercício)



Exercício 1

2. Suponha que duas empresas desejam empregá-lo e após considerar as vantagens de cada uma, você vai escolher aquela que lhe pagar melhor. Após certa pesquisa, você consegue a distribuição de salário das empresas, dadas segundo os gráficos abaixo.



Com base nas informações de cada gráfico, qual seria sua decisão?

Exercício 2

11. Vinte e uma pacientes de uma clínica médica tiveram o seu nível de potássio no plasma medido. Os resultados foram os seguintes:

Nível	freqüência
2,25 ⊢ 2,55	1
2,55 ⊢ 2,75	3
2,75 ⊢ 2,95	2
2,95 ⊢ 3,15	4
3,15 ⊢ 3,35	5
3,35 ⊢ 3,65	6

a. Construa o histograma.

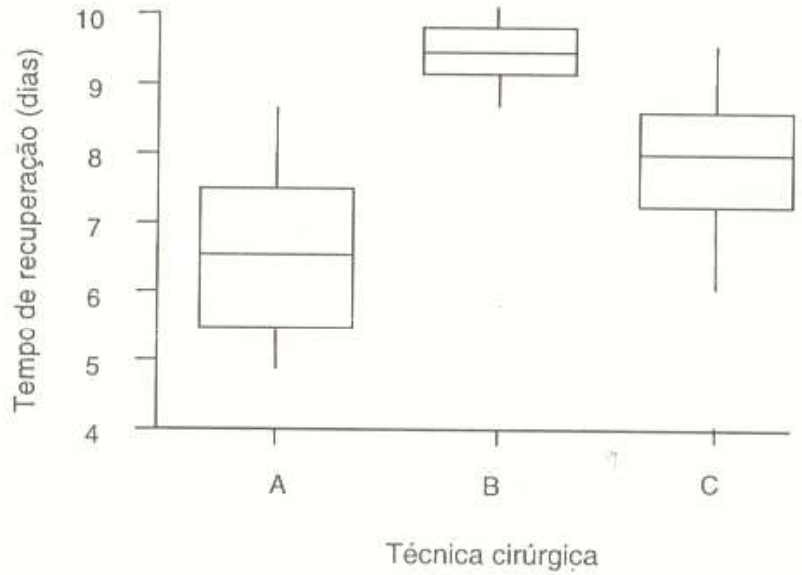
b. Determine os 1º, 2º e 3º quartis.

c. Qual a porcentagem dos valores que estão acima do nível 3?

d. Encontre a média e o desvio-padrão.

Exercício 3

22. Deseja-se comparar três técnicas cirúrgicas para a extração de dente de siso. Cada uma das técnicas foi aplicada em 20 pacientes e os resultados são apresentados a seguir.



a. Encontre valores aproximados para a mediana de cada técnica.
b. O *intervalo interquartil* é definido como a diferença entre o terceiro e o primeiro quartis. Calcule seu valor para cada uma das técnicas e comente.
c. Discuta a variabilidade do tempo de recuperação em cada técnica.
d. Se você é otimista, qual técnica escolheria?