Probabilidade

Análise Exploratória de Dados

Introdução

Tabelas Estatísticas

População, Amostra e Variáveis

Gráficos e Distribuição de Frequências

Renata Souza

Centro de Informática UFPE

Conceitos Antigos de Estatística

1) Simples contagem aritmética

Exemplos:

- Estatística de asfaltos, mais de 2000 acidentes em seis meses no Estado do Rio de Janeiro.
- O Estado do Ceará tem 679 indústrias.
- A população do Brasil no ano de 2008 é de 183.987.291.

2) Sinônimo de dados publicados oficialmente

- Publicações tais como: Anuário Estatístico do Brasil, Revista
- Brasileira de Estatística, IBGE, Boletim Estatístico.

Conceitos Antigos de Estatística

3) Simples transformações numéricas (percentagens, médias e razões, etc.)

Exemplos:

- Só 35 em 1000 alunos do curso primário concluem o Secundário.
- 58% dos veículos que rodam no país são nacionais.
- Um carro para 16 pessoas sem são Paulo.

Conceitos Antigos de Estatística

4) Construção de tabelas e gráficos

As informações contidas na tabela são compreendidas apenas avaliando o conteúdo da tabela.

Dados específicos são encontrados cruzando visualmente linhas e colunas.

Intenções de votos de candidatos por mês:

Candidato	Janeiro	Fevereiro	Março	Abril
João	3900	5600	3500	2300
Carlos	4500	5900	3100	3000
José	2100	4700	4000	3600

Tabelas Estatísticas

As tabelas devem obedecer à Resolução nº 886, de 26 de outubro de 1966, do Conselho Nacional de Estatística.

Cabeçalho, Rodapé e Corpo

Cabeçalho: Fornece uma breve descrição dos fins a que se destina

Rodapé: Fonte dos dados

Corpo: Contém os registros dos dados

Tabelas Estatísticas

Cabeçalho — Vendas no 1° Bimestre de 1996 da ABC Veículos

Corpo

Período	Unidades Vendidas
Janeiro/2008	20
Fevereiro/2008	10
Total	30

Fonte: ABC Veículos

Rodapé

Séries Estatísticas

É qualquer tabela que apresenta a distribuição de um conjunto de dados estatísticos em função da época, local ou espécie. Podem ser:

1. Série Temporal ou Cronológica;
2. Série Geográfica ou Histórica;
3. Série Específica (Categórica);
4. Distribuições de Frequências.

1. Série Temporal ou Cronológica

Identifica-se pelo caráter variável do fator cronológico. O local e a espécie são elementos fixos.

Ex.:

Nível pluviométrico por mês em Recife

Período	Nível (mm)
Janeiro/2008	142
Fevereiro/2008	274
Total Bimestral	416

Fonte: Embrapa

2. Série Geográfica ou Histórica

Apresenta como fator variável o fator geográfico. Também chamada de espacial, territorial ou de localização.

Média de habitantes por m^2 nas capitais Caracas, São Paulo e Recife em 2008

Período	Número
Caracas	1,42
São Paulo	2,50
Recife	2,10

Fonte: IBGE

3. Série Específica (Categórica)

O caráter variável é apenas o fato ou a espécie.

Número de títulos pernambucanos conquistados pelos principais times de Pernambuco

Time	Número
Sport	37
Náutico	21
Santa Cruz	24
Total	**82**

Fonte: FPF

4. Distribuições de Freqüências

Tabela onde os valores da variável não aparecem individualmente, mas agrupados em classes.

Intervalo: equivalente a [0;20[

Notas dos alunos do 2º período de Estatística em 2008

Notas	Número de Alunos
0 |-- 20	2
20 |-- 40	7
40 |-- 60	23
60 |-- 80	16
80 |-- 100	3
Total	**51**

Fonte: SIG@

População e Amostra

População

Conjunto de elementos que têm, em comum, determinada característica. As populações podem ser finitas ou infinitas. Além disso existem populações que, embora finitas, são consideradas infinitas para qualquer finalidade prática.

Amostra

Qualquer conjunto de elementos retirado da população, desde que esse conjunto seja não vazio e tenha um menor número de elementos que a população.

Esquema

Inferências Estatísticas:

Estimação de quantidades,

Exploração dos resultados,

Testes de Hipóteses

População e Amostra

A seleção da amostra pode ser feita e diversas maneiras dependentes entre outros fatores, do grau de conhecimento que temos da população e de recursos disponíveis.

A ideia é que amostra tenta fornecer um subconjunto de valores o mais parecido possível com a população que lhe dá origem.

A amostragem mais usada é a casual simples, em que selecionamos ao acaso, com ou sem reposição, os itens da população que farão parte da amostra.

Exemplo

Uma fração de fumantes preferem a marca de cigarros “Fumacê”. Aqueles que foram entrevistados constituem uma amostra representativa de todos os fumantes (que apesar de numericamente ser uma população finita, pode ser considerada infinita para efeitos práticos) .

Exemplos de tipos de Amostragem

1. Amostragem Aleatória:

Cada elemento da amostra é retirado aleatoriamente de toda a população (com ou sem reposição). Assim, cada possível amostra tem a mesma probabilidade de ser recolhida.

Ex.: Um professor deseja oferecer prêmios (5 livros) aos seus alunos em número de 35 e resolve apelar para um sorteio.

Exemplos de tipos de Amostragem

2. Amostragem Estratificada:

Subdividir a população em pelo menos dois grupos distintos que partilham alguma característica e, em seguida, recolher uma amostra de cada um dos grupos (ou estratos).

Ex.: A turma tem 13 alunos e 23 alunas.

A amostra é $\frac{5}{35} = \frac{1}{7}$

(1/7) de 13 = 1,86 ≈ 2 (1/7) de 23 = 3,14 ≈ 3

Exemplos de tipos de Amostragem

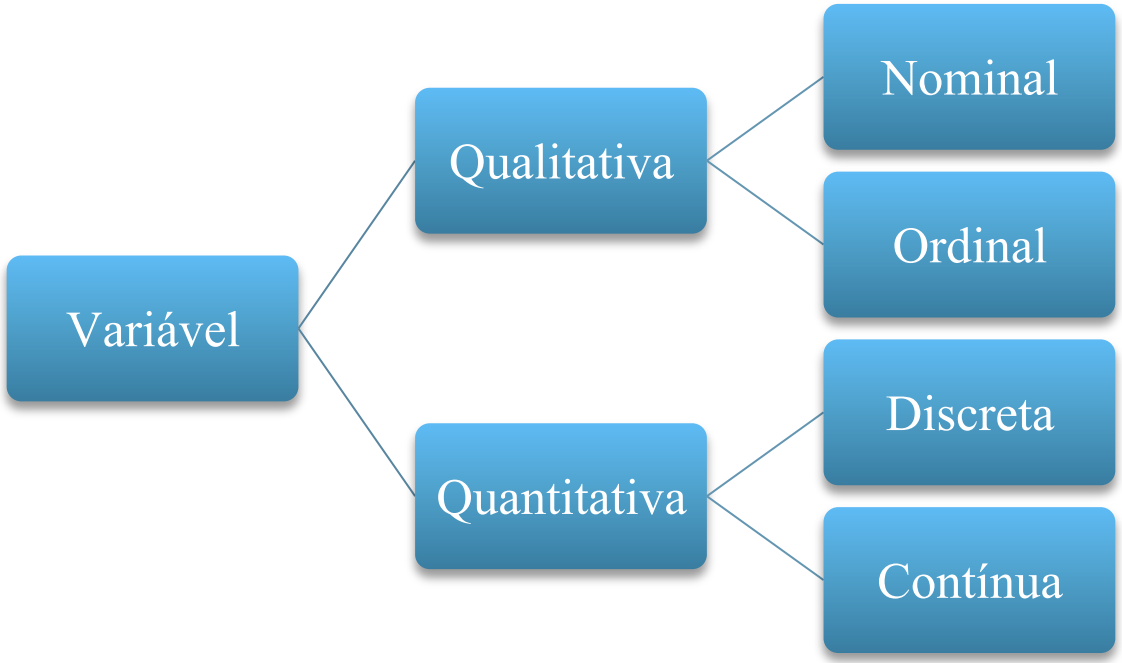
3. Amostragem Sistemática:

Quando os elementos da população se apresentam ordenados e a retirada dos elementos da amostra é feita periodicamente, temos uma amostragem sistemática.

Ex.: Sorteia-se um número x (0 < x < 50) e faz r = 50/5 = 10 para encontrar qual dos cinco alunos, numerados de 0 a 4, vão apresentar o trabalho.

Variável

Característica que pode ser observada (ou mensurada) nos elementos da população, devendo ter pelo menos um resultado para cada elemento observado.



Variável

1. Qualitativa: O resultado da variável é um atributo ou uma qualidade.

1.1. Qualitativa Ordinal: representam com uma ordenação natural.

Ex.: Classe social: A- alta, C- média, D- baixa

Escolaridade: 1- Primária, 2- Secundária, 3- Superior

1.2. Qualitativa Nominal: não existe ordenação dentre as categorias

Ex.: sexo, cor dos olhos, fumante/não fumante, doente/sadio

Variável

2. Quantitativa: O resultado é um número numa escala pré-determinada.

2.1 Discreta: Os resultados possíveis são números inteiros. Ex.: números de alunos.

2.2 Contínua: O resultado está em um intervalo dos números reais.

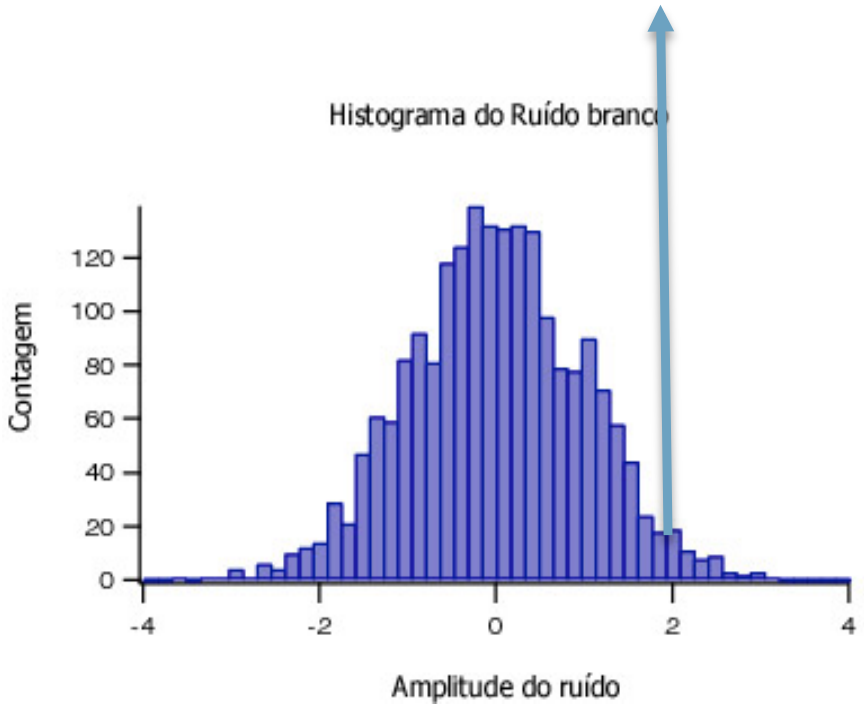
Ex.: atraso de transmissão de bytes por uma rede de internet.

Histogramas

Representação gráfica de uma distribuição de frequências por meio de retângulos justapostos.

A distribuição de frequência é o método mais útil para descrever resultados obtidos com respeito a uma variável.

Na amostra existem, aproximadamente, 20 elementos com amplitude de ruído igual a 2.



Distribuição de Frequência

Tabela onde os valores da variável não aparecem individualmente, mas agrupados em classes.

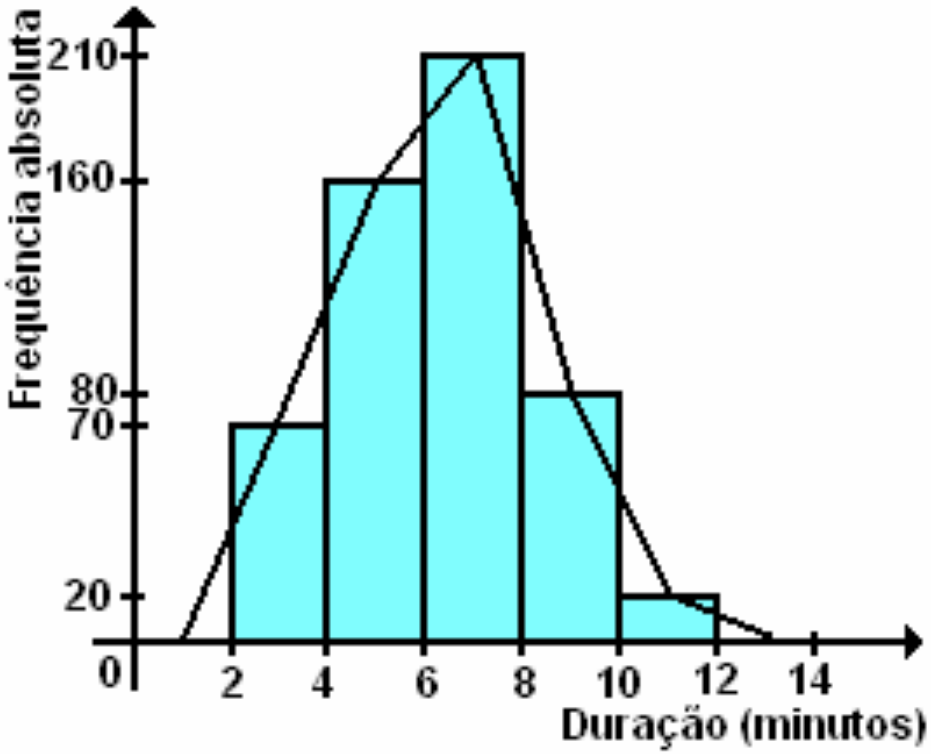
Com *muitos intervalos* corremos o risco de não realçar os aspectos relevantes;

Mas com *poucos intervalos*, os grupos se tornam muito abrangentes, impedindo uma maior precisão;

Importante: definir a amplitude dos intervalos.

Polígono de Frequências

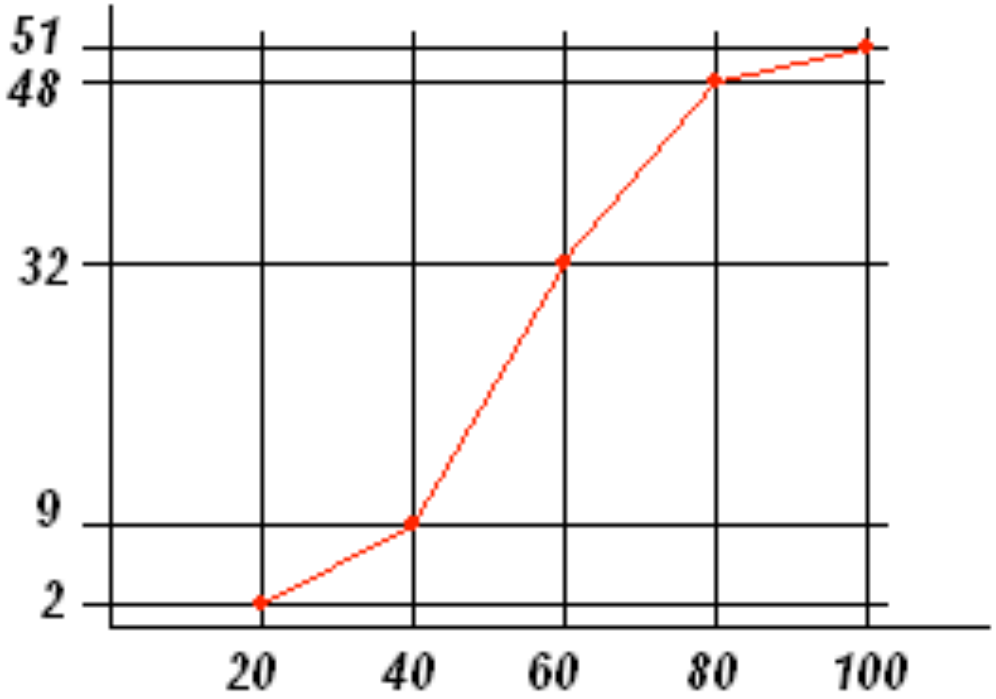
É um gráfico de linha, sendo as frequências os pontos médios dos intervalos das classes.



Polígono de Frequência Acumulada

Um ponto no gráfico representa a soma de todas as frequências das classes anteriores mais a que esse ponto corresponde.

Notas	Nº de Alunos
0 |-- 20	2
20 |-- 40	7
40 |-- 60	23
60 |-- 80	16
80 |-- 100	3
Total	**51**



Gráficos

Representam os resultados obtidos, permitindo chegar-se a conclusões sobre a evolução de fenômeno ou sobre como se relacionam os valores da série;

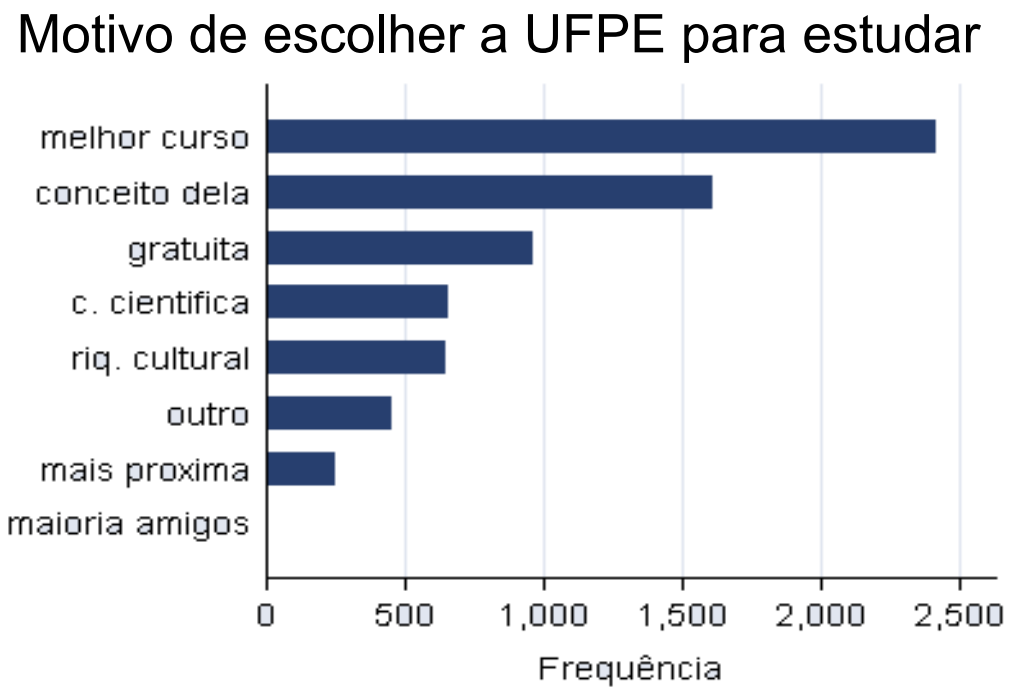
Dependendo do critério de quem irá fazer o gráfico, as séries podem ser representadas por:

1. Gráfico de Barras;
2. Gráfico de Colunas;
3. Gráfico de Setor;
4. Gráfico de Hastes.

1. Gráfico de Barras

Representação gráfica da distribuição de frequência para variáveis Qualitativas;

As barras são espaçadas, possuem a mesma largura e são dispostas horizontalmente.



2. Gráfico de Colunas

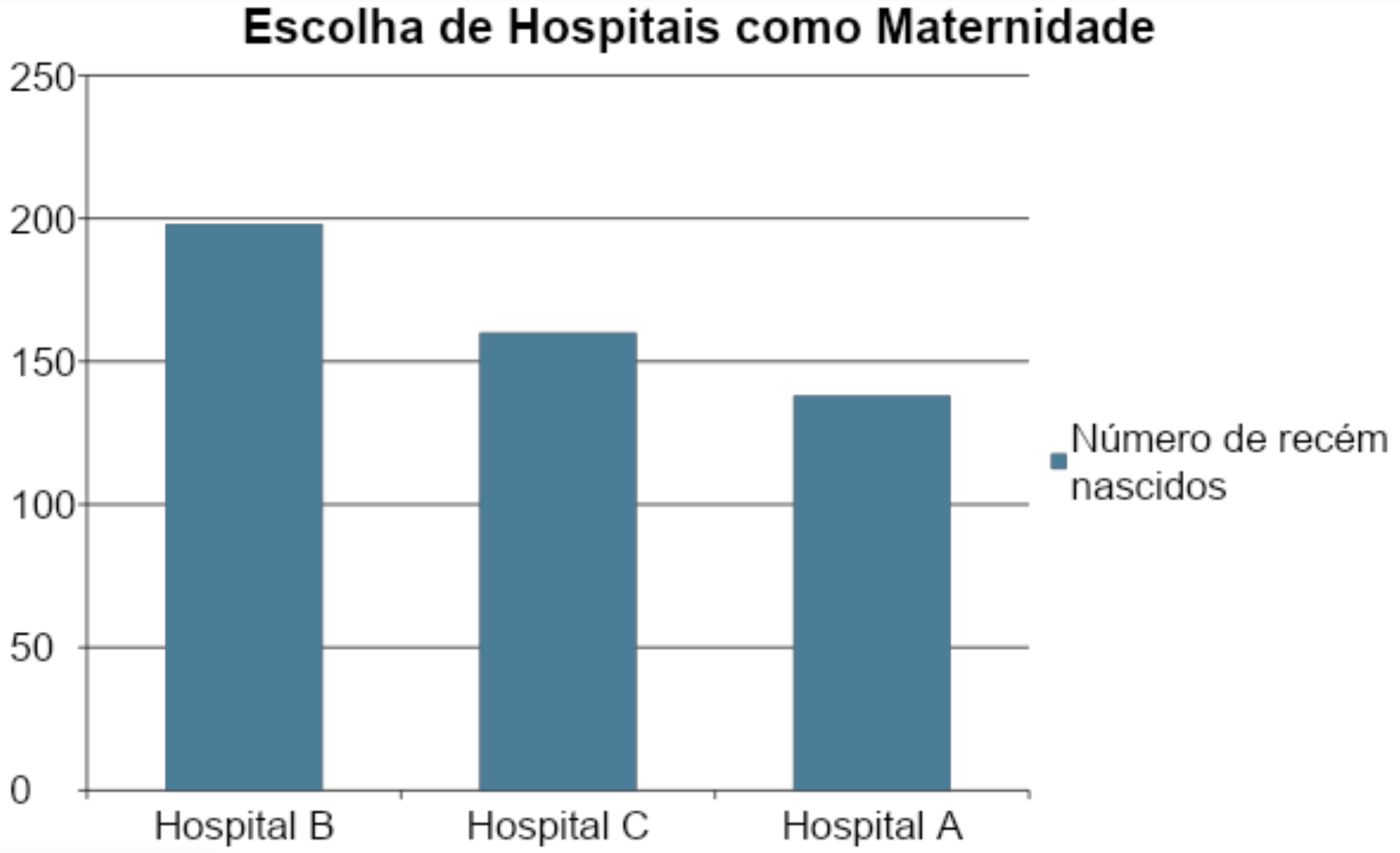
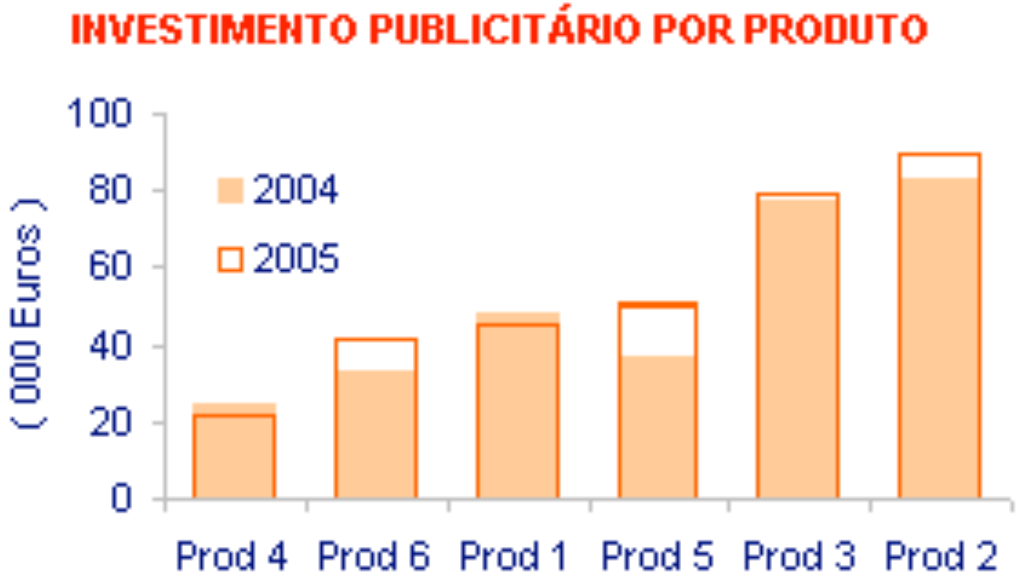


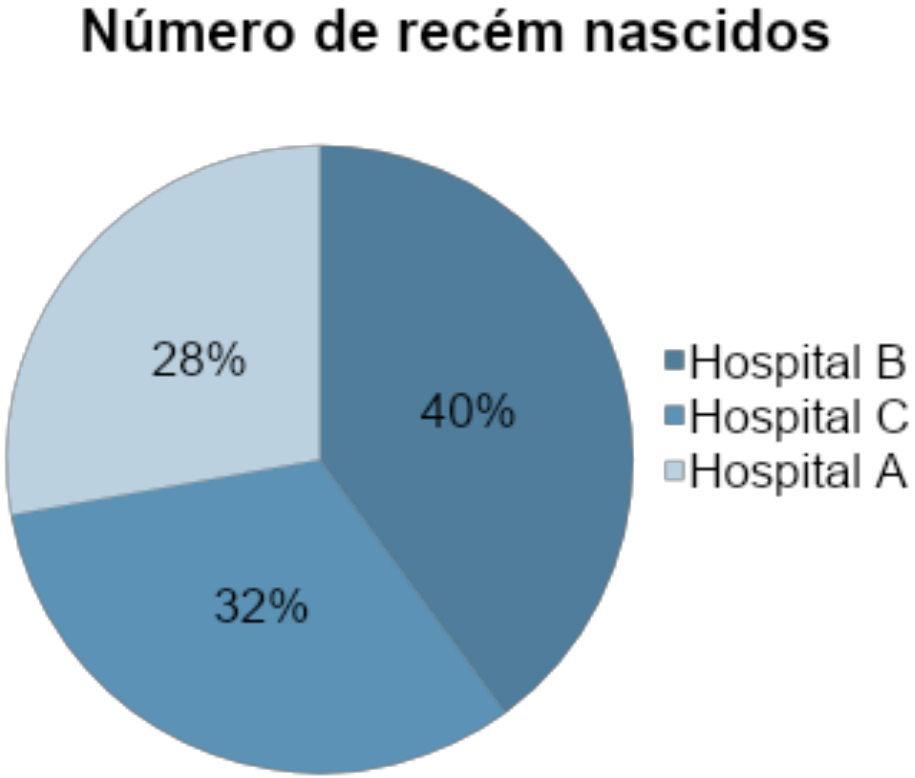
Gráfico de Colunas

Os gráficos de coluna são úteis para mostrar alterações de dados em um período de tempo ou para ilustrar comparações entre itens.



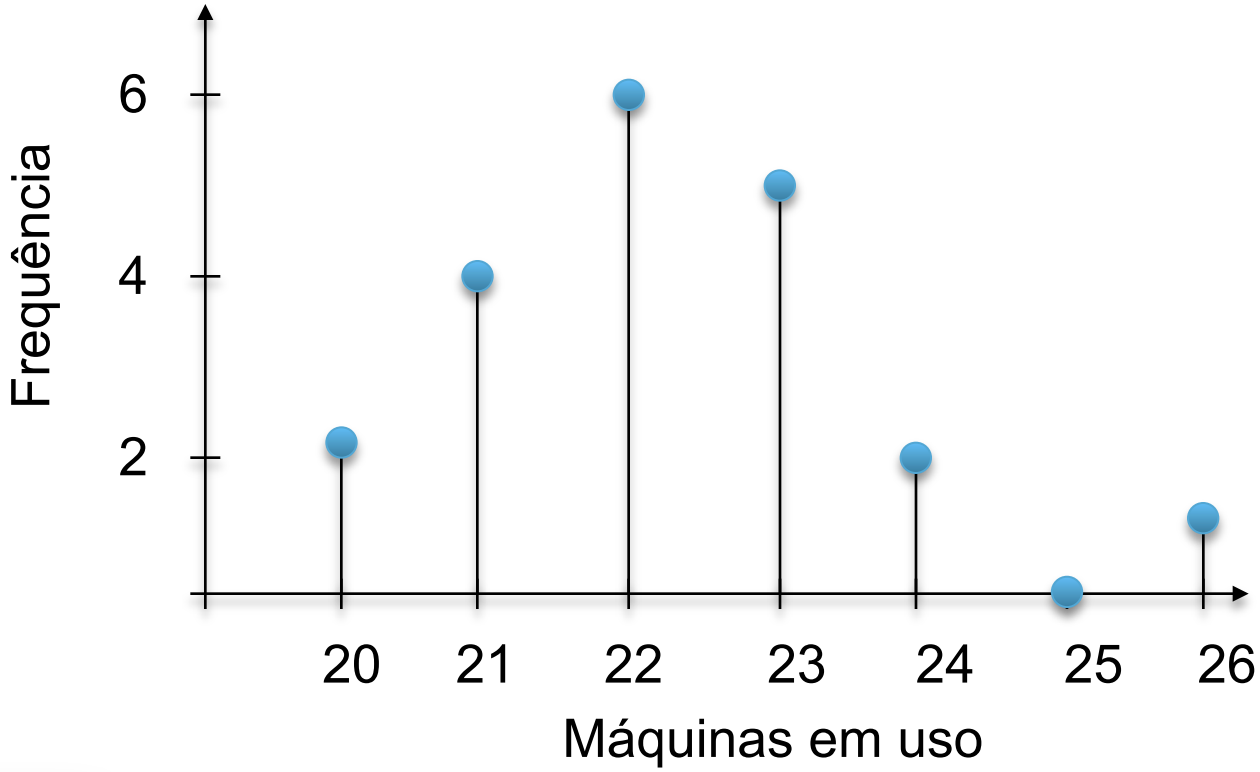
3. Gráfico de Setor

O gráfico de setores é usado para mostrar a importância relativa das proporções. Então esse gráfico trabalha com porcentagens.



4. Gráfico de Hastes

Esse tipo de gráfico é útil na representação de variáveis de tempo discreto



Construção de tabelas de distribuição de frequência

Objetivo: construir tabelas de distribuição de frequência a partir de dados brutos (n observações).

1º Passo: determinar a amplitude total;

2º Passo: estimar o número de intervalos;

3º Passo: estimar a amplitude dos intervalos;

4º Passo: esquematizar a tabela de acordo com as informações dos passos anteriores.

Exemplo

Tempo em segundos para carga de um aplicativo num sistema compartilhado (50 observações):

5,2	6,4	5,7	8,3	7,0	5,4	4,8	9,1	5,5	6,2	4,9	5,7	6,3
5,1	8,4	6,2	8,9	7,3	5,4	4,8	5,6	6,8	5,0	6,7	8,2	7,1
4,9	5,0	8,2	9,9	5,4	5,6	5,7	6,2	4,9	5,1	6,0	4,7	18,1
5,3	4,9	5,0	5,7	6,3	6,0	6,8	7,3	6,9	6,5	5,9		

1º Passo: Determinar a amplitude total (range)

5,2	6,4	5,7	8,3	7,0	5,4	4,8	9,1	5,5	6,2	4,9	5,7	6,3
5,1	8,4	6,2	8,9	7,3	5,4	4,8	5,6	6,8	5,0	6,7	8,2	7,1
4,9	5,0	8,2	9,9	5,4	5,6	5,7	6,2	4,9	5,1	6,0	**4,7**	**18,1**
5,3	4,9	5,0	5,7	6,3	6,0	6,8	7,3	6,9	6,5	5,9		

Menor tempo

Maior tempo

Amplitude total R = 18,1 – 4,7 = 13,4

2º Passo: estimar o nº de intervalos (classes)

- O número de intervalos $K = \sqrt{n}$, para $n > 25$
  - $K = 5$, para $n < 25$;
  - $K = \sqrt{50} = 7{,}07$

- Ou pode usar a fórmula de *Sturges* $K = 1 + 3{,}22 \log n$
  - $K = 1 + 3{,}22 \log 50 = 7$
  - n é o tamanho da amostra.

- Lembrar da importância sobre o número dos intervalos

3º Passo: estimar a amplitude dos intervalos

- Amplitude dos intervalos

$$h = \frac{R}{K}$$

$$h = \frac{13{,}4}{7} = 1{,}914 = 1{,}92$$

4º Passo: montar a tabela

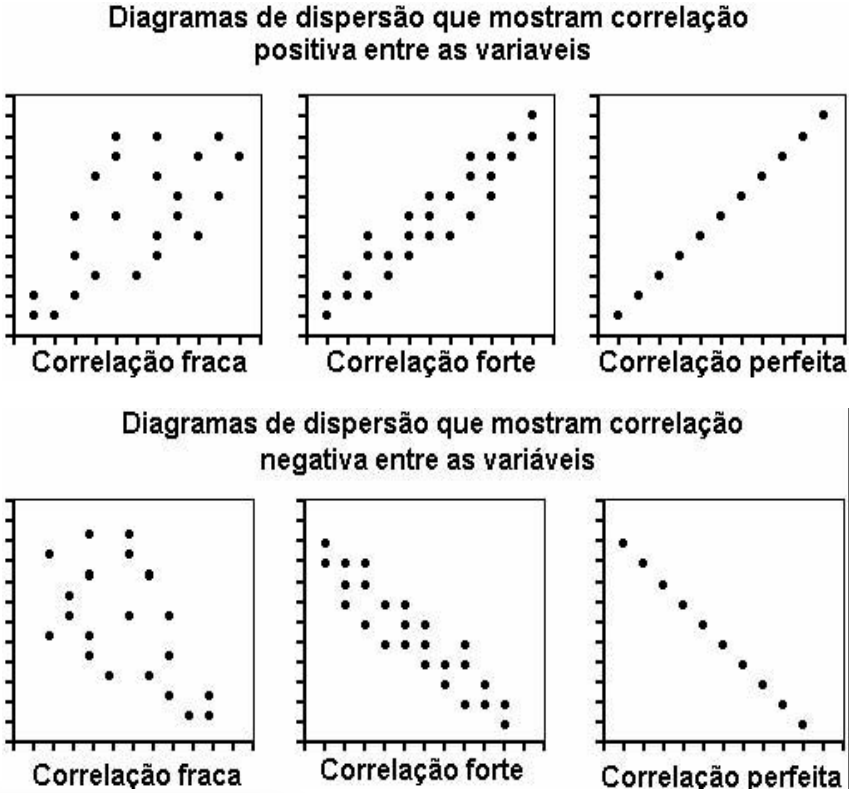
4,70 + h

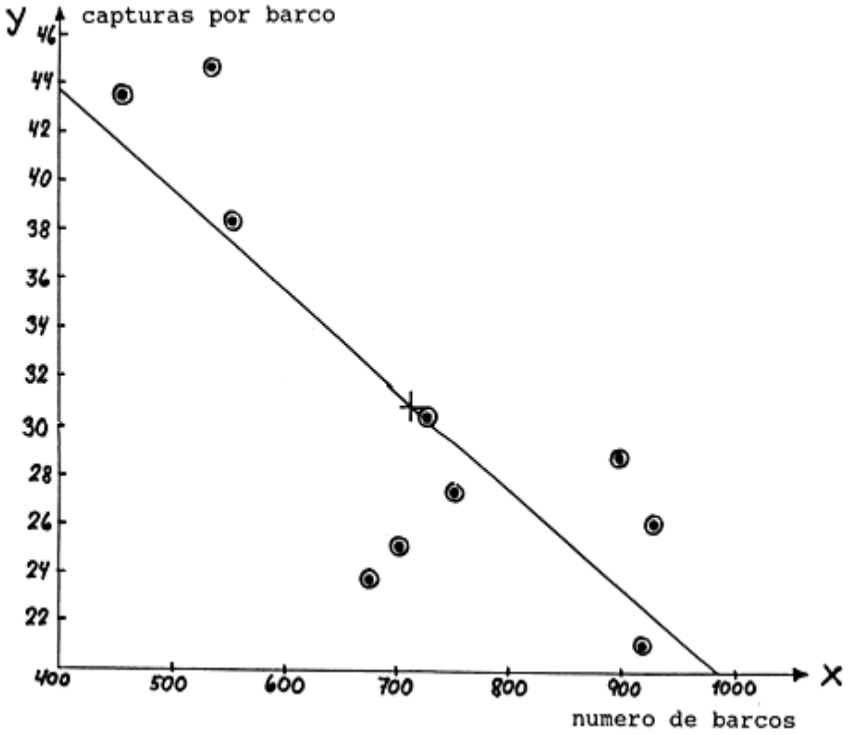
Valor mínimo

Tempo	Frequência absoluta	Frequência relativa
4,70 |-- 6,62	34	68%
6,62 |-- 8,54	12	24%
8,54 |-- 10,46	3	6%
10,46 |-- 12,38	0	0%
12,38 |-- 14,30	0	0%
14,30 |-- 16,22	0	0%
16,22 |-- 18,14	1	2%
Total	**50**	**100%**

Diagramas de Dispersão

Serve para saber se existe alguma correlação (forte, fraca, moderada, positiva, negativa) entre duas variáveis.

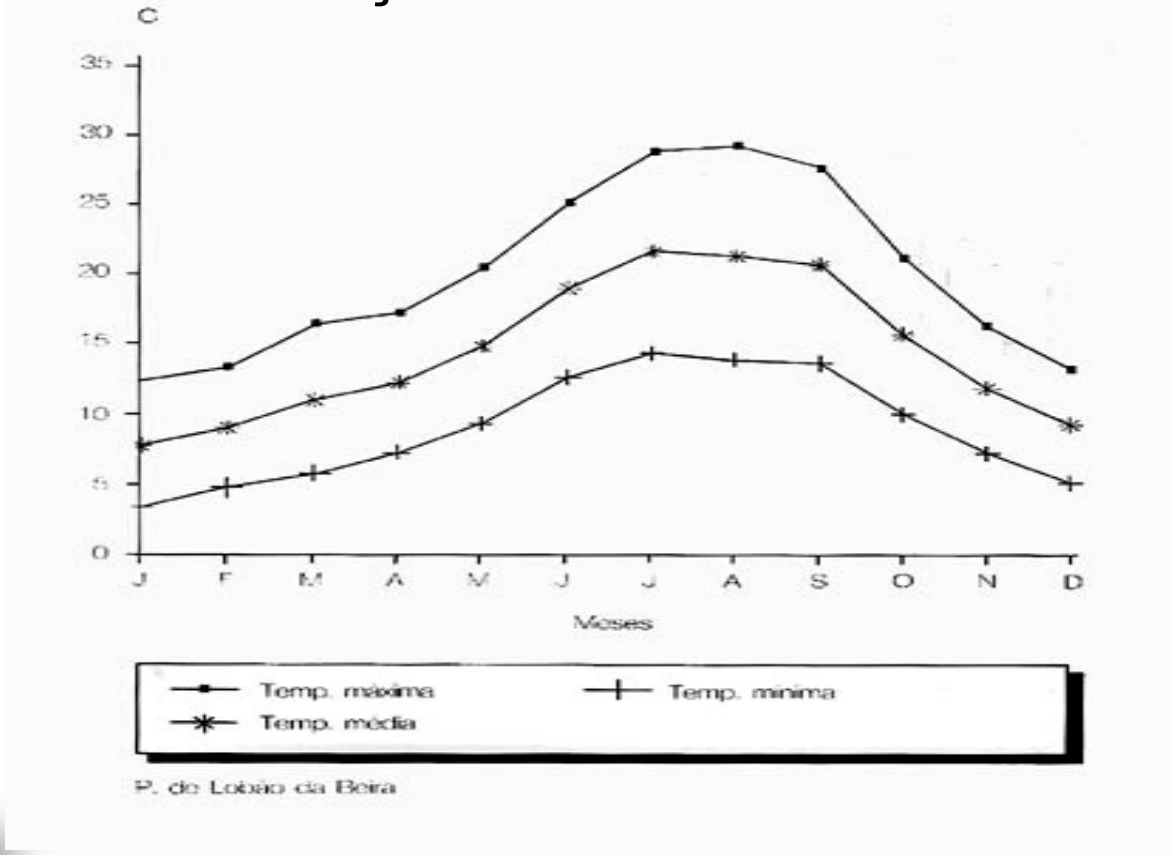




Centro de Informática UFPE

Gráficos de Curvas

Usados em processos para se acompanhar a evolução de uma variável em relação a um ou mais limites existentes.



Considerações

Gráficos setoriais são particularmente úteis para visualizar diferenças entre classes. Eles não acomodam grandes quantidades de categorias.

Nesse caso:

- reagrupar as menos importantes em um grupo chamado outros ou,
- utilizar um gráfico de barras, sendo que estas devem vir separadas;

Considerações

Tipo de variável ou série	Método mais usado ou adequado	Comentário
Dados qualitativos	Gráfico de barras, colunas ou circulares (tipo torta)	
Variáveis discretas	Medidas intervalares. Gráfico de hastes	
Variáveis contínuas	Gráficos em forma de histogramas e polígonos de frequência	O uso de polígonos de frequência induz o leitor a aceitar a continuidade da variável apresentada.
Séries cronológicas	Gráfico de colunas, curvas ou barras	
Séries específicas e geográficas	Gráfico de colunas, barra ou setor	O gráfico tipo setor permite uma maior visualização das partes frente do todo.

Exercício

Dada a amostra:

3,2 - 4,1 - 4,9 - 5,0 - 7,3 - 6,7 - 6,6 - 7,4 - 7,1 - 4,0 - 5,5 - 5,4 - 6,5 - 6,5 - 7,1 - 5,2 - 8,3 - 5,7 - 6,8- 6,4

Pede-se:

a) Construir a distribuição de frequência;
b) Construir o gráfico das frequências;
c) Determinar as frequências relativas;
d) Determinar as frequências acumuladas;
e) Qual a amplitude amostral e de cada classe;
f) Qual a porcentagem de elementos maiores que 5;
g) Construir o histograma.