Bases Computacionais da Ciência

Noções de Estatística, Distribuições, e Gráficos de Dispersão

Prof. Ronaldo Cristiano Prati

Bloco A, sala 513-2

ronaldo.prati@ufabc.edu.br

Objetivo

- Introduzir algumas ferramentas básicas de Análise Estatística:

> Permitem visualizar e compreender características de dados experimentais e realizar formas simples de inferência.

- Familiarizar o aluno com o uso da ferramenta para **automatizar tarefas de análise estatística**, que seria por demais tediosas ou (difíceis) de se realizar manualmente.
- A objetivo desta aula NÃO é esgotar o assunto da análise estatística de dados

Motivação

A Estatística é um ramo da Matemática que estuda como se pode usar uma amostra para tirar conclusões sobre um universo maior de objetos, levando em conta que sempre há variação e incerteza nas medidas consideradas.

A Estatística está presente na base de toda a ciência experimental, pois ela fornece diretrizes para a coleta de dados, permite comparar diferentes hipóteses e avaliar a acurácia dos resultados obtidos experimentalmente.

Estatística

Conjunto de técnicas que permite de forma sistemática as seguintes operações sobre dados:

- Organizar
- Descrever
- Analisar
- Interpretar

Análise estatística

A Análise Estatística pode ser dividida em duas áreas:

- Estatística descritiva
- Estatística indutiva / inferencial

Estatística Descritiva

Voltada a apresentação, organização e resumo numérico dos dados:

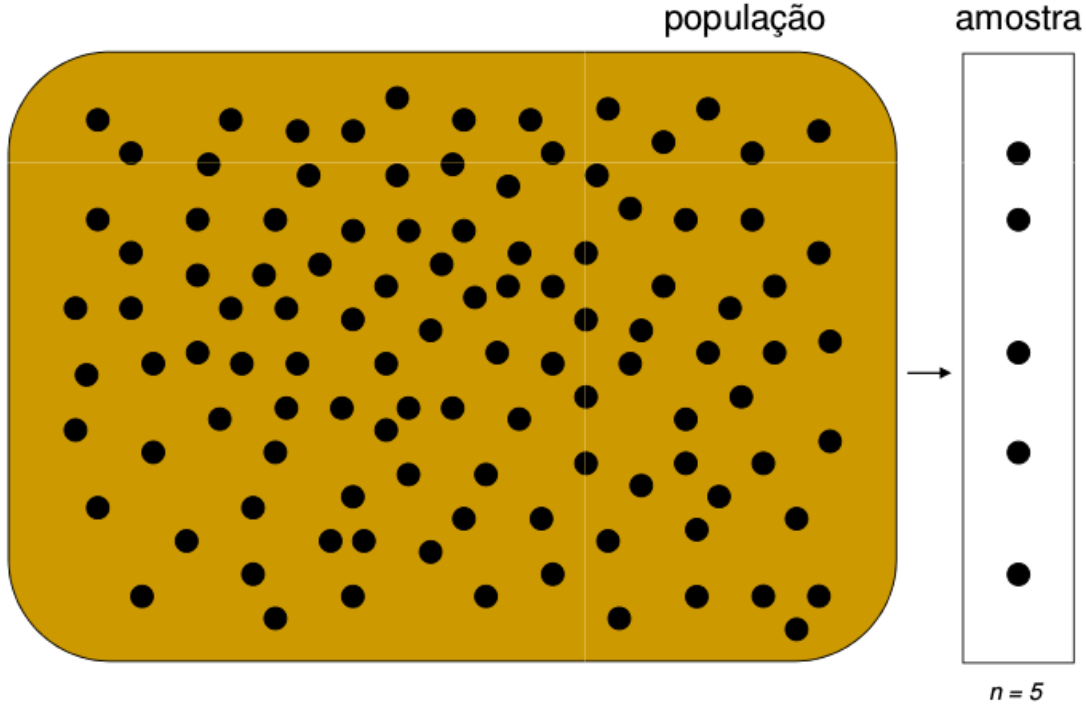
- Pode incluir a construção de gráficos, gráficos tabelas e computação de várias medidas, tais como, medidas de tendência central (ex. a média), de dispersão (ex. a variância), de frequência (ex. percentagem), etc.
- O propósito deste tipo de estatística é de fazer com que os dados coletados sejam compreendidos mais facilmente, seja em forma gráfica ou numérica (tabelas).

Cuidado: "estatística" é o termo para o conjunto de procedimentos que conhecemos como "a estatística" mas também o termo geral para medidas descritivas deste tipo – p.ex., a média é "uma estatística".

Estatística indutiva/inferencial

Voltada a realizar estimativas a partir de uma amostra ou testar ideias teóricas (hipóteses) com dados experimentais

> Se uma amostra é representativa de uma população, conclusões importantes sobre a população podem ser inferidas de sua análise .



Análise estatística: exemplos

Estatística Descritiva:

- O número de acidentes (= frequência) nas rodovias federais no estado de São Paulo antes e depois da Lei Seca;
- Gráfico com a distribuição da idade dos ingressantes nos bacharelados interdisciplinares da UFABC.

Estatística Indutiva/Inferencial:

- Estimação da porcentagem da população que votará para um/a determinado/a candidato/a à presidência, junto com uma margem de erro (“intervalo de confiança”);
- Teste estatístico de tendência de queda nas populações de atum-rabilho entre 2000 e 2010, a partir de observações sistemáticas

Variáveis

Medição de certas características de interesse para cada um dos casos presentes na amostra.
As características medidas são conhecidas como variáveis. Por exemplo:

- Estudo sobre habitantes de uma cidade, as variáveis podem ser: Altura, sexo, cor do cabelo, cor dos olhos, etc

Divididas em dois tipos:

- Independente
- Dependente

Tipos de variáveis

Independente:

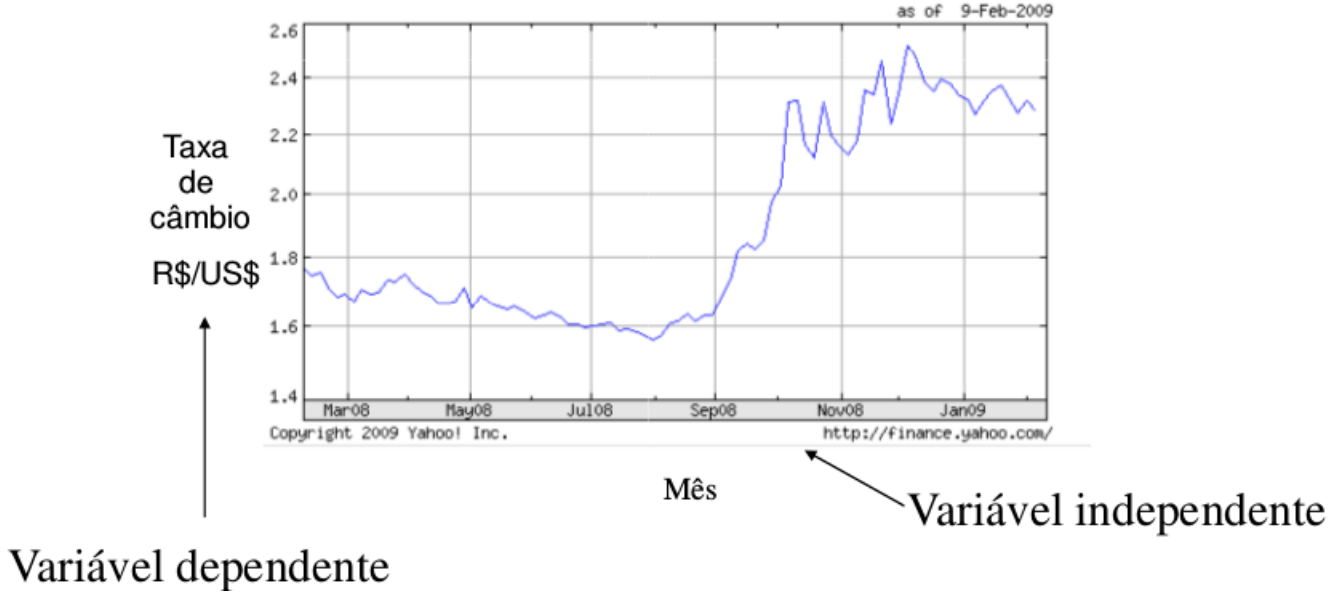
- Valores manipulados ou selecionados pelo pesquisador (meio, idade, mês).
- Podem ser ou não a "causa" da variável dependente.

Dependente:

- Valores observados, contados, medidos, ... que não estejam sob controle direto do pesquisador (velocidade, taxa de câmbio).
- Podem ser "causadas" ou não pela variável independente.

Tipos de variáveis

- Quando não há relação causal óbvia entre duas ou mais variáveis, qual é ‘independente ‘ ou ‘dependente’ é uma questão de rótulo.
- A variável ‘dependente’ é a que que analisamos em função dos valores de uma outra variável.



Variáveis discretas e contínuas

- Variáveis quantitativas: expressadas em valores numéricos (qualitativas)
- Discretas: Conjunto enumerável de valores
    - Nominais (categóricas) sem ordem natural de valores: {presente, ausente}, {homem, mulher}, estado de origem (UF), base DNA A/C/T/G.
    - Ordinais: com ordem natural de valores: Classe sócio-econômica (A-E ou “baixa”, “média”, “alta”), avaliação em escala Likert (nota 1-5), {PP, P, M, G, GG}, número de acidentes.

Variáveis discretas e contínuas

- Contínuas: Conjunto não-enumerável, valores reais, não discretizados
    - Grandezas físicas ou químicas: Velocidade, força, probabilidade, concentração, acidez, taxa de câmbio.

Medidas de tendência central

É conveniente dispor de medidas que informem sobre a amostra de maneira mais resumida do que os dados brutos são capazes de fazer.

As medidas de tendência central cumprem este papel, dando o valor do ponto em torno do qual os dados se distribuem:
Valor ‘médio’ ou ‘típico’ de um conjunto de dados.

Por exemplo, são medidas de tendência central:

- Média
- Mediana
- Moda

Média aritmética

É o ‘centro de gravidade’ dos dados.
Soma de um conjunto de valores dividida pelo número de valores do conjunto:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Em que:

- N é o número total de observações
- X_i é um valore do conjunto

Média aritmética

Para calcular a média aritmética no Scilab:

```
// lista de 1 a 10
X1 = 1:1:10;
mean(X)
//calculando pela fórmula
sum(X)/length(X)

// lista de 10 números aleatórios
X2 = rand(1:10);
mean(X2)

// lista de 1 milhão de números aleatórios
X3 = rand(1:1000000);
mean(X3)
```

Mediana

Valor central do conjunto que divide a distribuição em duas partes iguais
(mesmo número de escores abaixo e acima do valor).

Os dados devem estar ordenados

Posição da mediana:

$$i = \frac{N+1}{2}$$

Mediana

Para calcular a média mediana no Scilab:

```
// lista de números
X = [3,5,6,4,5,8,9,6,2,7,5];
median(X)
// posição do elemento da mediana
i = (length(X)+1)/2
// elmento na posição da mediana
gsort(X)(i)
```

Mediana

Para calcular a média mediana no Scilab:

```
// lista de números
X = [3,5,6,4,5,8,9,6,2,7,5,100];

media(X)

gsort(X)
```

Por que isso ocorre?

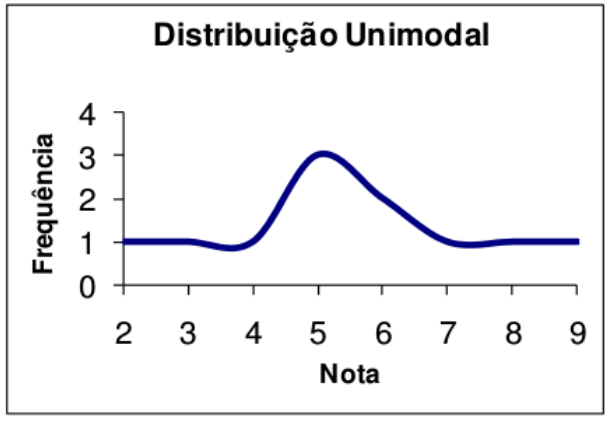
- No caso de um número par de sujeitos a mediana é a média entre os dois valores centrais.

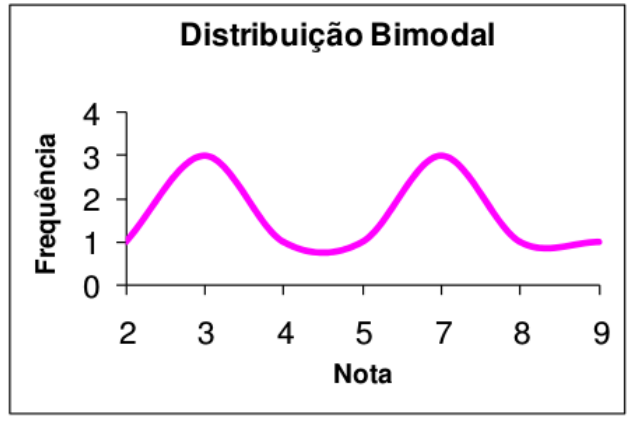
Moda

É a categoria que ocorre com maior frequência.

- A moda pode não existir OU pode não ser única.
- Exemplos:
  1,1,3,3,5,7,7,7,11,13 ➔ moda 7
  3,5,8,11,13,18 ➔ não tem moda
  3,5,5,5,6,6,7,7,7,11,12 ➔ tem duas modas: 5,7 (bimodal).

Sujeitos	Notas
2	1
3	3
4	1
5	1
7	3
8	1
9	1





Scilab: Histograma

```
//lista de números
x = [3,5,6,4,5,8,9,6,2,7,5]

//cria um histograma com a frequência dos valores de x
histplot(10,x)

//cria um histograma com a frequência dos valores de x
histplot(10,x,style=12,polygon=%t)

// cria uma lista com 1 milhão de números aleatórios
x = rand(1:1000000)

//cria um histograma com a frequência dos valores de x
hisplot(10,x)
```

Medidas de dispersão

- O processo de trabalhar com amostras introduz uma variabilidade dos resultados obtidos, pois cada amostra vai ter características ligeiramente diferentes
- Essa variabilidade afeta nosso grau de confiança nos resultados. Por isso, as medidas de variabilidade (ou dispersão) têm papel central na Estatística.

Dentre as medidas de dispersão tem-se:

- Variância
- Desvio-padrão

Variância

É a 'Média' dos quadrados dos desvios, onde desvio é a diferença entre cada dado e a média do conjunto.

Dados (X)	Desvios $(X-\overline{X})$	Quadrados dos Desvios $(X-\overline{X})^2$
0	-5	25
4	-1	1
6	1	1
8	3	9
7	2	4
$\overline{X}=5$	$\sum(X-\overline{X})=0$	$\sum(X-\overline{X})^2=40$

$$s^2 = \frac{\sum(X-\bar{X})^2}{N-1} = \frac{40}{4} = 10$$

Desvio padrão

Raiz quadrada da variância

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}} = \sqrt{10} \approx 3,16$$

Variância e Desvio padrão

Para se calcular a variância e o desvio padrão no scilab:

```
x = [0,4,6,8,7]

// calculo da variância
variance(x)

//calculo do desvio padrão
stdev(x)
```

Representação gráfica das informações

Gráficos tem por finalidade representar os resultados obtidos.
Permite chegar a conclusões sobre a evolução do fenômeno ou sobre como se relacionam os valores.
Não há uma única maneira de representar graficamente uma série estatística.
Escolha do gráfico mais apropriado ficará a critério do analista.
Alguns critérios:

- Simplicidade
- Clareza
- Veracidade

Gráfico de colunas

O **gráfico de barras** é um gráfico com barras retangulares e comprimento proporcional aos valores que ele representa. As barras podem ser desenhadas verticalmente ou horizontalmente. O gráfico de barras vertical as vezes é chamado de **gráfico de colunas**.

```
x = [1:10];
n = [8, 6, 13, 10, 6, 4, 16, 7, 8, 5];

clf();
bar(x,n);
plot(x,n,'r*-');
```

Gráficos de dispersão

Os gráficos de dispersão são representações de dados de duas (ou mais) variáveis que são exibidos como uma coleção de pontos, cada um com o valor de uma variável determinando a posição no eixo horizontal e o valor da outra variável determinando a posição no eixo vertical.

Opcionalmente, se adiciona uma "linha de tendência" a respeito desses dados (mais na próxima aula).

Gráficos de dispersão

```
// primeira variável
m = [1,2,3,4,5];

// segunda variável
g = [300,430,700,1200,2300];

//gráfico de dispersão
plot(s,g,'bo')
```

Atividade Semana 2

O objetivo desta aula é comparar o uso de uma ferramenta baseada no conceito de planilha eletrônica (Calc, equivalente ao Excel) com o Scilab
Office Calc

- Faça o download da planilha "aula2_dados_doenca_coronariana.xls" e abra com o Calc
- Faça as atividades 2 e 3, conforme descrita na apostila (páginas 100 a 106)
- Acrescente o gráfico de disperção (Seção 3.3.7, página 88 a 95 da apostila) de tabaco com alcool
-

Atividade Semana 2

Vamos fazer a mesma atividade com o scilab

Navegue na aba esquerda até a pasta onde a planilha está localizada. Para carregar a planilha no Scilab, use os comandos

```
// lê a planilha
planilha = readxls("aula02_dados_doenca_coronariana.xls");

// carrega os dados da primeira planilha
dados = planilha(1);

// ignora o cabeçalho
dados = dados(2:41,:);
```

Atividade Semana 2

Calculando as estatísticas com o Scilab

```
// seleciona a coluna da idade é (coluna 2)
idade = dados(:,2);

// imprime as estatísticas
printf("média: %f\n",mean(idade));
printf("variância: %f\n",variance(idade));
printf("desvio padrao: %f\n",stdev(idade));
printf("mediana: %f\n",median(idade));
```

Gráficos

```
histplot(10,idade) // cria um histograma de idade

tabaco = dados(:,4)'   //'converte de coluna para linha
alcool = dados(:,6)'   //'converte de coluna para linha

// calcula a linha de tendência (mais na próxima aula)
[a,b] = reglin(tabaco,alcool);

plot(tabaco,alcool,'ro'); // gráfico de dispersão
plot([0,20],[b,a*20+b]); // plota a linha de regressão nos
                // (0,b) e (20,a*20+b)
```

Atividade Semana 2

Entregar a planilha com as atividades 2 e um documento com os comandos/gráficos gerados no scilab até o dia 18/06