

# ***APOSTILA***

# ***ESTATÍSTICA***

---



*Luis Felipe Dias Lopes, Dr.*  
[lflopes@smail.ufsm.br](mailto:lflopes@smail.ufsm.br), [phil.zaz@zaz.com.br](mailto:phil.zaz@zaz.com.br)

DE - UFSM  
2003

# Sumário

---

## **1 Conceitos básicos**

- 1.1 População x Amostra
- 1.2 Censo x Amostragem
- 1.3 Dado x Variável
- 1.4 Parâmetros x estatísticas
- 1.5 Arredondamento de dados
- 1.6 Fases do método estatístico

## **2 Representação tabular**

- 2.1 Representação esquemática
- 2.2 Elementos de uma tabela
- 2.3 Séries estatísticas
- 2.4 Distribuição de frequência

## **3 Representação gráfica**

- 3.1 Gráficos de Linhas
- 3.2 Gráficos de colunas ou barras
- 3.3 Gráficos circulares ou de Setores (Pie Charts)
- 3.4 Gráfico Pictorial - Pictograma
- 3.5 Gráfico Polar
- 3.6 Cartograma
- 3.7 Gráficos utilizados para a análise de uma distribuição de frequência

## **4 Medidas descritivas**

- 4.1 Medidas de posição
- 4.2 Medidas de variabilidade ou dispersão
- 4.3 Medidas de dispersão relativas
- 4.4 Momentos, assimetria e curtose
- 4.5 Exercícios

## **5 Probabilidade e variáveis aleatórias**

- 5.1 Modelos matemáticos
- 5.2 Conceitos em probabilidade
- 5.3 Conceitos de probabilidade
- 5.4 Exercícios
- 5.5 Teorema de Bayes
- 5.6 Variáveis aleatórias
- 5.7 Função de probabilidade

5.8 Exemplos

5.9 Exercícios

## **6 Distribuições de Probabilidade**

6.1 Distribuições discretas de probabilidade

6.2 Exercícios

6.2 Distribuições contínuas de probabilidade

6.4 Exercícios

## **7 Amostragem**

7.1 Conceitos em amostragem

7.2 Planos de amostragem

6.3 Tipos de amostragem

7.4 Amostragem com e sem reposição

7.5 Representação de uma distribuição amostral

7.6 Distribuições amostrais de probabilidade

7.7 Exercícios

7.8 Estatísticas amostrais

7.9 Tamanho da amostra

## **8 Estimação de parâmetros**

8.1 Estimação pontual

8.2 Estimação intervalar

8.3 Exercícios

## **9 Testes de hipóteses**

9.1 Principais conceitos

8.2 Teste de significância

9.3 Exercícios

9.4 Testes do Qui-quadrado

9.5 Exercícios

## **10 Regressão e Correlação**

10.1 Introdução

10.2 Definição

10.3 Modelo de Regressão

10.4 Método para estimação dos parâmetros  $\alpha$  e  $\beta$

10.5 Decomposição da variância Total

10.6 Análise de Variância da Regressão

10.7 Coeficiente de Determinação ( $r^2$ )

10.8 Coeficiente de Correlação ( $r$ )

10.9 Exercícios

## **11 Referências bibliográficas**

# 1 Conceitos Básicos

## 1.1 População x Amostra

- **População (N):** Conjunto de todos os elementos relativos a um determinado fenômeno que possuem pelo menos uma característica em comum, a população é o conjunto Universo, podendo ser finita ou infinita.
- **Finita** - apresenta um número limitado de observações, que é passível de contagem.
- **Infinita** - apresenta um número ilimitado de observações que é impossível de contar e geralmente esta associada a processos.
- **Amostra (n):** É um subconjunto da população e deverá ser considerada finita, a amostra deve ser selecionada seguindo certas regras e deve ser representativa, de modo que ela represente todas as características da população como se fosse uma fotografia desta.



Uma população pode, mediante processos operacionais, ser considerada infinita, pois a mesma irá depender do tamanho da amostra. Se a frequência relativa entre amostra e população for menor do que 5% ela é considerada infinita, se a frequência relativa for maior do 5% ela é considerada finita.

## 1.2 Censo x Amostragem

- **Pesquisa Estatística:** É qualquer informação retirada de uma população ou amostra, podendo ser através de Censo ou Amostragem.
- **Censo:** É a coleta exaustiva de informações das "N" unidades populacionais.
- **Amostragem:** É o processo de retirada de informações dos "n" elementos amostrais, no qual deve seguir um método criterioso e adequado (tipos de amostragem).

## 1.3 Dado x Variável

- **Dados estatísticos:** é qualquer característica que possa ser observada ou medida de alguma maneira. As matérias-primas da estatística são os dados observáveis.
- **Variável:** É aquilo que se deseja observar para se tirar algum tipo de conclusão, geralmente as variáveis para estudo são selecionadas por processos de amostragem. Os símbolos utilizados para representar as variáveis são as letras maiúsculas do alfabeto, tais como X, Y, Z, ... que pode assumir qualquer valor de um conjunto de dados. As variáveis podem ser classificadas dos seguintes modos:

- **Qualitativas (ou atributos):** São características de uma população que não pode ser medidas.

**Nominal :** são utilizados símbolos, ou números, para representar determinado tipo de dados, mostrando, assim, a qual grupo ou categoria eles pertencem.

**Ordinal ou por postos:** quando uma classificação for dividida em categorias ordenadas em graus convencionados, havendo uma relação entre as categorias do tipo “maior do que”, “menor do que”, “igual a”, os dados por postos consistem de valores relativos atribuídos para denotar a ordem de primeiro, segundo, terceiro e, assim, sucessivamente.

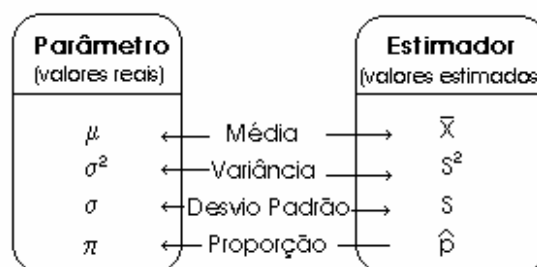
- **Quantitativas:** São características populacionais que podem ser quantificadas, sendo classificadas em discretas e contínuas.

**Discretas:** são aquelas variáveis que pode assumir somente valores inteiros num conjunto de valores. É gerada pelo processo de **contagem**, como o número de veículos que passa em um posto de gasolina, o número de estudantes nesta sala de aula.

**Contínuas:** são aquelas variáveis que podem assumir um valor dentro de um intervalo de valores. É gerada pelo processo de **medição**. Neste caso serve como exemplo o volume de água em um reservatório ou o peso de um pacote de cereal.

#### 1.4 Parâmetros x Estatísticas

- **Parâmetros:** são medidas populacionais quando se investiga a população em sua totalidade, neste caso é impossível fazer inferências, pois toda a população foi investigada.
- **Estatísticas ou Estimadores:** são medidas obtidas da amostra, torna-se possível neste caso utilizarmos as teorias inferências para que possamos fazer conclusões sobre a população.



## 1.5 Arredondamento de Dados

Regras: Portaria 36 de 06/07/1965 - INPM  $\Rightarrow$  Instituto Nacional de Pesos e Medidas.

1ª) Se o primeiro algarismo após aquele que formos arredondar for de 0 a 4, conservamos o algarismo a ser arredondado e desprezamos os seguintes.

Ex.: 7,34856 (para décimos)  $\rightarrow$  7,3

2ª) Se o primeiro algarismo após aquele que formos arredondar for de 6 a 9, acrescenta-se uma unidade no algarismo a ser arredondado e desprezamos os seguintes.

Ex.: 1,2734 (para décimos)  $\rightarrow$  1,3

3ª) Se o primeiro algarismo após aquele que formos arredondar for 5, seguido apenas de zeros, conservamos o algarismo se ele for **par** ou aumentamos uma unidade se ele for **ímpar**, desprezando os seguintes.

Ex.: 6,2500 (para décimos)  $\rightarrow$  6,2

12,350 (para décimos)  $\rightarrow$  12,4



Se o 5 for seguido de outros algarismos dos quais, pelo menos um é diferente de zero, aumentamos uma unidade no algarismo e desprezamos os seguintes.

Ex.: 8,2502 (para décimos)  $\rightarrow$  8,3

8,4503 (para décimos)  $\rightarrow$  8,5

4ª) Quando, arredondarmos uma série de parcelas, e a soma ficar alterada, devemos fazer um novo arredondamento (por falta ou por excesso), na maior parcela do conjunto, de modo que a soma fique inalterada.

Ex.:  $17,4\% + 18,4\% + 12,3\% + 29,7\% + 22,2\% = 100\%$

arredondando para inteiro:

$17\% + 18\% + 12\% + 30\% + 22\% = 99\%$

$17\% + 18\% + 12\% + 31\% + 22\% = 100\%$

## 1.6 Fases do método estatístico

O método estatístico abrange as seguintes fases:

### **a) Definição do Problema**

Consiste na:

- formulação correta do problema;
- examinar outros levantamentos realizados no mesmo campo (revisão da literatura);
- saber exatamente o que se pretende pesquisar definindo o problema corretamente (variáveis, população, hipóteses, etc.)

### **b) Planejamento**

Determinar o procedimento necessário para resolver o problema:

- Como levantar informações;
- Tipos de levantamentos: Por Censo (completo);  
Por Amostragem (parcial).
- Cronograma, Custos, etc.

### **c) Coleta ou levantamento dos dados**

Consiste na obtenção dos dados referentes ao trabalho que desejamos fazer.

A coleta pode ser: Direta - diretamente da fonte;  
Indireta - feita através de outras fontes.

Os dados podem ser obtidos pela própria pessoa (primários) ou se baseia no registro de terceiros (secundários).

### **d) Apuração dos Dados ou sumarização**

Consiste em resumir os dados, através de uma contagem e agrupamento. É um trabalho de coordenação e de tabulação.

Apuração: manual, mecânica, eletrônica e eletromecânica.

### **e) Apresentação dos dados**

É a fase em que vamos mostrar os resultados obtidos na coleta e na organização.

Esta apresentação pode ser: Tabular (apresentação numérica)  
Gráfica (apresentação geométrica)

### **f) Análise e interpretação dos dados**

É a fase mais importante e também a mais delicada. Tira conclusões que auxiliam o pesquisador a resolver seu problema.

## 2 Representação tabular

Consiste em dispor os dados em linhas e colunas distribuídas de modo ordenado. A elaboração de tabelas obedece à Resolução nº 886, de 26 de outubro de 1966, do Conselho Nacional de Estatística. As normas de apresentação são editadas pela Fundação Brasileira de Geografia e Estatística (IBGE).

### 2.1 Representação esquemática

Título

=====  
Cabeçalho  
=====

Corpo

-----  
Rodapé

### 2.2 Elementos de uma tabela

- **Título:** O título deve responder as seguintes questões:
  - O que? (Assunto a ser representado (Fato));
  - Onde? (O lugar onde ocorreu o fenômeno (local));
  - Quando? (A época em que se verificou o fenômeno (tempo)).
- **Cabeçalho:** parte da tabela na qual é designada a natureza do conteúdo de cada coluna.
- **Corpo:** parte da tabela composta por linhas e colunas.
- **Linhas:** parte do corpo que contém uma seqüência horizontal de informações.
- **Colunas:** parte do corpo que contém uma seqüência vertical de informações.
- **Coluna Indicadora:** coluna que contém as discriminações correspondentes aos valores distribuídos pelas colunas numéricas.
- **Casa ou célula:** parte da tabela formada pelo cruzamento de uma linha com uma coluna.
- **Rodapé:** É o espaço aproveitado em seguida ao fecho da tabela, onde são colocadas as notas de natureza informativa (fonte, notas e chamadas).
- **Fonte:** refere-se à entidade que organizou ou forneceu os dados expostos.
- **Notas e Chamadas:** são esclarecimentos contidos na tabela (**nota** - conceituação geral; **chamada** - esclarecer minúcias em relação a uma célula).



### 2.3 Séries Estatísticas

Uma série estatística é um conjunto de dados ordenados segundo uma característica comum, as quais servirão posteriormente para se fazer análises e inferências.

- **Série Temporal ou Cronológica:** É a série cujos dados estão dispostos em correspondência com o tempo, ou seja, varia o tempo e permanece constante o fato e o local.

Produção de Petróleo Bruto no Brasil de 1976 a 1980 (x 1000 m<sup>3</sup>)

Anos	Produção
1976	9 702
1977	9 332
1978	9 304
1979	9 608
1980	10 562

Fonte: Conjuntura Econômica (fev. 1983)

- **Série Geográfica ou Territorial:** É a série cujos dados estão dispostos em correspondência com o local, ou seja, varia o local e permanece constante a época e o fato.

População Urbana do Brasil em 1980 (x 1000)

Região	População
Norte	3 037
Nordeste	17 568
Sudeste	42 810
Sul	11 878
Centro-Oeste	5 115
Total	80 408

Fonte: Anuário Estatístico (1984)

- **Série Específica ou Qualitativa:** É a série cujos dados estão dispostos em correspondência com a espécie ou qualidade, ou seja, varia o fato e permanece constante a época e o local.

População Urbana e Rural do Brasil em 1980 (x 1000)

Localização	População
Urbana	80 408
Rural	38 566
Total	118 974

Fonte: Anuário Estatístico (1984)

- **Série Mista ou Composta:** A combinação entre duas ou mais séries constituem novas séries denominadas compostas e apresentadas em tabelas de dupla entrada. O nome da série mista surge de acordo com a combinação de pelo menos dois elementos.

Local + Época = Série Geográfica Temporal

População Urbana do Brasil por Região de 1940 a 1980 (x 1000)

Anos	REGIÕES				
	N	NE	SE	S	CO
1940	406	3 381	7 232	1 591	271
1950	581	4 745	10 721	2 313	424
1960	958	7 517	17 461	4 361	1 007
1970	1 624	11 753	28 965	7 303	2 437
1980	3 037	17 567	42 810	11 878	5 115

Fonte: Anuário Estatístico (1984)

## 2.4 Distribuição de Frequência

É o tipo de série estatística na qual permanece constante o fato, o local e a época. Os dados são colocados em classes preestabelecidas, registrando a frequência de ocorrência. Uma distribuição de frequência pode ser classificada em discreta e intervalar.

**a) Distribuição de Frequência Discreta ou Pontual:** É uma série de dados agrupados na qual o número de observações está relacionado com um ponto real.

Notas do Aluno "X" na Disciplina de Estatística segundo critérios de avaliação do DE da UFSM – 1990

$X_i$	$f_i$
6.3	2
8.4	3
5.3	2
9.5	3
6.5	5
$\Sigma$	15

Fonte: Departamento de Estatística (1990)

**b) Distribuição de Frequências Intervalar:** Na distribuição de frequência, os intervalos parciais deverão ser apresentados de maneira a evitar dúvidas quanto à classe a que permanece determinado elemento.

O tipo de intervalo mais usado é do tipo fechado a esquerda e aberto a direita, representado pelo símbolo: |---.

Altura em centímetros de 160 alunos do Curso de Administração da UFSM - 1990

Altura (cm)		$X_i$	$f_i$
150	--- 158	154	18
158	--- 166	162	25
166	--- 174	170	20
174	--- 182	178	52
182	--- 190	186	30
190	--- 198	194	15
$\Sigma$		----	160

Fonte: Departamento de Estatística (1990)

### Elementos de uma Distribuição de Frequências:

➤ **Classe ou Classe de Frequência (K):** É cada subintervalo (linha) na qual dividimos o fenômeno.

Para determinar o número de classes a partir dos dados não tabelados, podemos usar a **Fórmula de Sturges**, mas deve-se saber que existem outros métodos de determinação do número de classes em uma tabela de frequência. O que se deseja fazer é apenas comprimir um conjunto de dados em uma tabela, para facilitar a visualização e interpretação dos mesmos.

$$n(K) = 1 + 3.3 \log n, \text{ onde "n" é nº de informações.}$$



Além da Regra de Sturges, existem outras fórmulas empíricas para resolver o problema para determinação do número de classes  $[n(k)]$ , há quem prefira  $n(k) \cong \sqrt{n}$ . Entretanto, a verdade é que essas fórmulas não nos levam a uma decisão final; esta vai depender na realidade de um julgamento pessoal, que deverá estar ligado a natureza dos dados, procurando, sempre que possível, evitar classes com frequências nulas ou frequências relativas exageradamente grandes.

➤ **Limite de Classe ( $l_i$  ou  $L_i$ ):** São os valores extremos de cada classe.

$l_i$  = limite inferior da i-ésima classe;

$L_i$  = limite superior da i-ésima classe;

- **Amplitude do intervalo de classe (h):** É a diferença entre dois limites inferiores ou superiores consecutivos.

$$h = l_n - l_{n-1} \quad \text{ou} \quad h = L_n - L_{n-1}$$



A amplitude do intervalo de classe deve ser constante em toda a distribuição de freqüências intervalar.

- **Amplitude total (H):** É a diferença entre o limite superior da última classe e o limite inferior da 1ª classe, ou a diferença entre último e o primeiro elemento de um conjunto de dados postos em ordem crescente.

$$H = L_n - l_1$$

- **Ponto médio de classe (X<sub>i</sub>):** É a média aritmética simples do limite inferior com o limite superior de uma mesma classe.

$$X_i = \frac{l_i + L_i}{2}$$

ou a partir do X<sub>1</sub> os demais pontos médios pode ser determinado por:

$$X_n = X_{n-1} + h$$



Quando substituirmos os intervalos de classes pelos pontos médios (X<sub>i</sub>), ter-se-á uma **distribuição de freqüência pontual**.

- **Freqüência absoluta (f<sub>i</sub>):** É a quantidade de valores em cada classe

$$n = \sum_{i=1}^n f_i = f_1 + f_2 + \dots + f_n$$

- **Freqüência Acumulada (F<sub>i</sub>):** É o somatório da freqüência absoluta da i-ésima classe com a freqüência absoluta das classes anteriores, ou a freqüência acumulada da classe anterior.

$$F_n = \sum_{i=1}^n f_i = n$$

- **Freqüência Relativa (fr<sub>i</sub>):** É o quociente entre a freqüência absoluta da i-ésima classe com o somatório das freqüências.

$$fr_i = \frac{f_i}{\sum_{i=1}^n f_i} \quad \text{Obs.:} \quad \sum_{i=1}^n fr_i = 1$$

- **Freqüência Relativa Acumulada (Fr<sub>i</sub>):** É o somatório da freqüência relativa da i-ésima classe com as freqüências relativas das classes anteriores.

$$Fr_n = \sum_{i=1}^n fr_i = 1$$

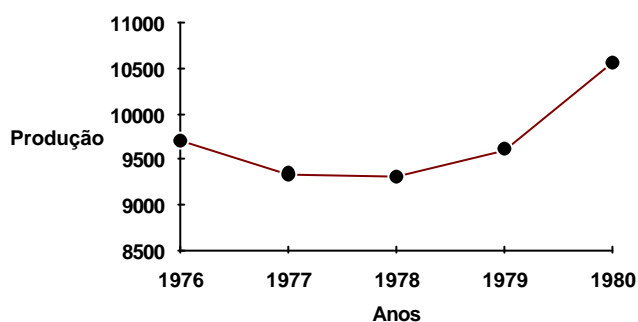
## 3 Representação gráfica

Os gráficos são uma forma de apresentação visual dos dados. Normalmente, contém menos informações que as tabelas, mas são de mais fácil leitura. O tipo de gráfico depende da variável em questão

### 3.1 Gráficos de Linhas

Usado para ilustrar uma série temporal.

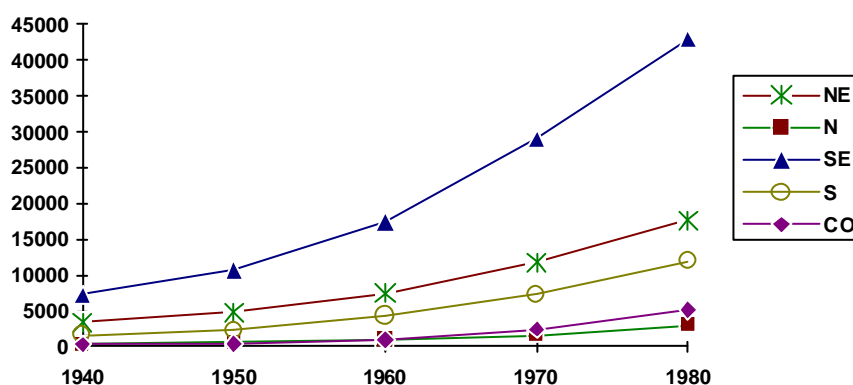
Produção de Petróleo Bruto no Brasil de 1976 a 1980 (x 1000 m<sup>3</sup>)



Fonte: Conjuntura Econômica (Fev. 1983)

#### 3.1.1 Gráfico de linhas comparativas

População Urbana do Brasil por Região de 1940 a 1980 (x 1000)



Fonte: Anuário Estatístico (1984)

### 3.2 Gráficos de colunas ou barras

Representação gráfica da distribuição de freqüências. Este gráfico é utilizado para variáveis nominais e ordinais.

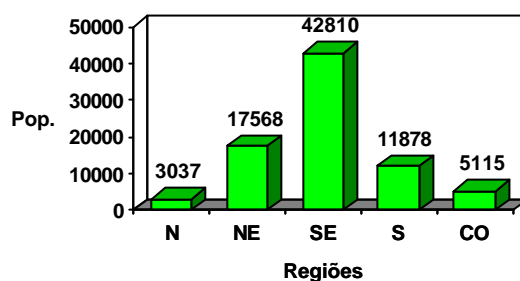
Características:

- todas as barras devem ter a mesma largura
- devem existir espaços entre as barras

#### 3.2.1 Gráfico de Colunas

Usado para ilustrar qualquer tipo de série.

População Urbana do Brasil em 1980 (x 1000)



Fonte: Anuário Estatístico (1984)

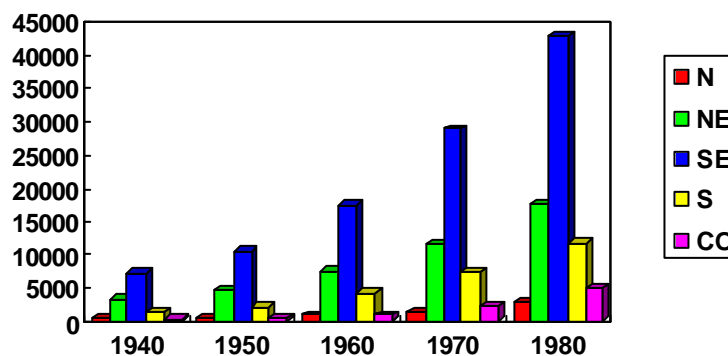


As larguras das barras que deverão ser todas iguais podendo ser adotado qualquer dimensão, desde que seja conveniente e desde que não se superponham. O número no topo de cada barra pode ou não omitido, se forem conservados, a escala vertical pode ser omitida.

#### 3.2.2.1 Gráfico de colunas comparativas

##### a) Colunas Justapostas (gráfico comparativo)

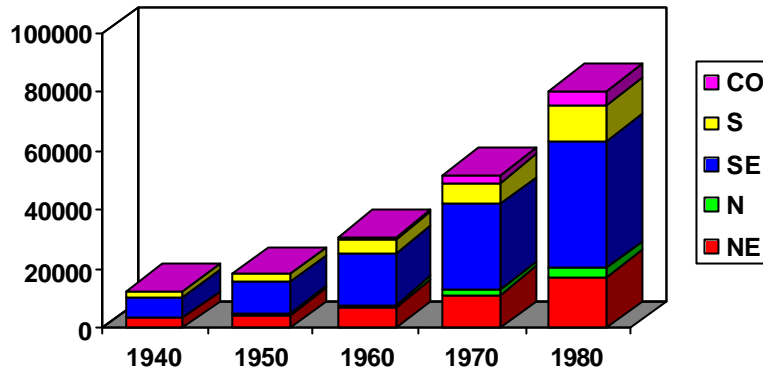
População Urbana do Brasil por Região de 1940 a 1980 (x 1000)



Fonte: Anuário Estatístico(1984)

### b) Colunas Sobrepostas (gráfico comparativo)

População Urbana do Brasil por Região de 1940 a 1980 (x 1000)

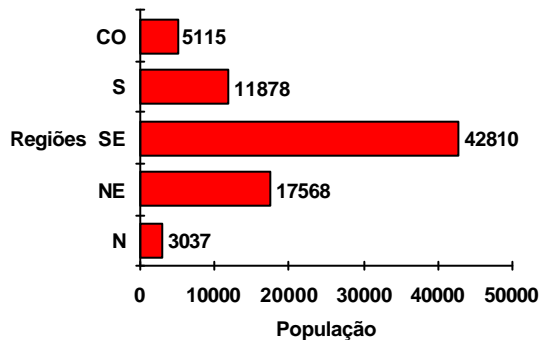


Fonte: Anuário Estatístico (1984)

### 3.2.2 Gráfico de Barras

As regras usadas para o gráfico de barras são iguais as usadas para o gráfico de colunas.

População Urbana do Brasil em 1980 (x 1000)



Fonte: Anuário Estatístico (1984)



Assim como os gráficos de Colunas podem ser construídos gráficos de barras comparativas.

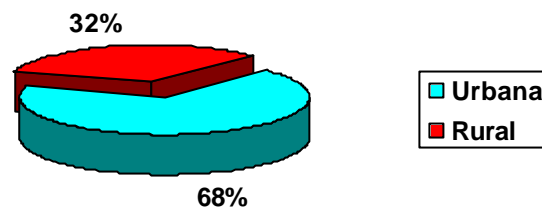
### 3.3 Gráficos circulares ou de Setores (Pie Charts)

Representação gráfica da frequência relativa (percentagem) de cada categoria da variável. Este gráfico é utilizado para variáveis nominais e ordinais. É uma opção ao gráfico de barras quando se pretende dar ênfase à comparação das percentagens de cada categoria. A construção do gráfico de setores segue uma regra de 3 simples, onde as frequências de cada classe correspondem ao ângulo que se deseja representar em relação a frequência total que representa o total de 360°.

*Características:*

- A área do gráfico equivale à totalidade de casos ( $360^\circ = 100\%$ );
- Cada “fatia” representa a percentagem de cada categoria

População Urbana e Rural do Brasil em 1980 (x 1000)

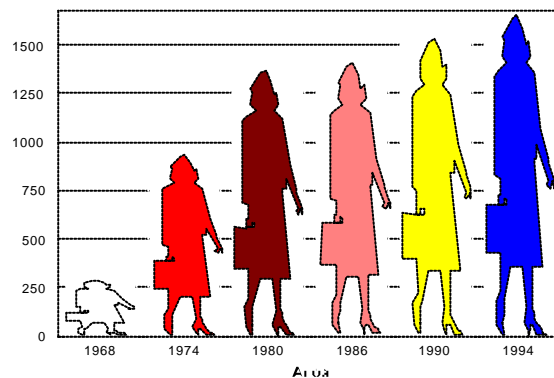


Fonte: Anuário Estatístico (1984)

### 3.4 Gráfico Pictorial - Pictograma

Tem por objetivo despertar a atenção do público em geral, muito desses gráficos apresentam grande dose de originalidade e de habilidade na arte de apresentação dos dados.

Evolução da matrícula no Ensino Superior no Brasil de 1968 a 1994 (x 1000)

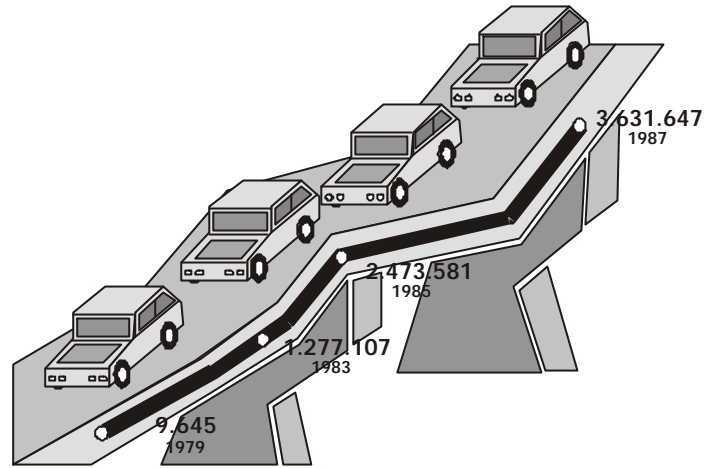


Fonte: Grandes números da educação brasileira março de 1996

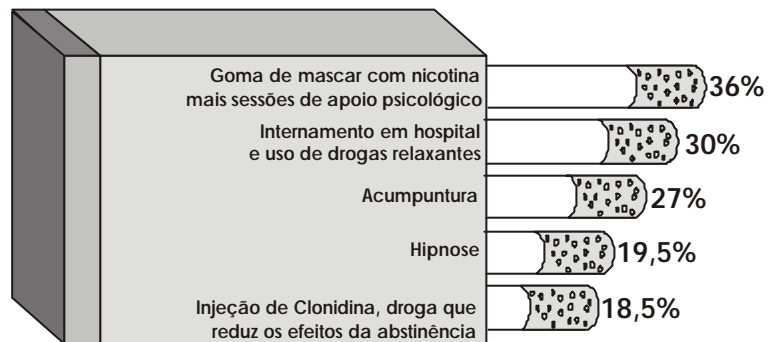


### 3.4.1 Exemplos de pictogramas

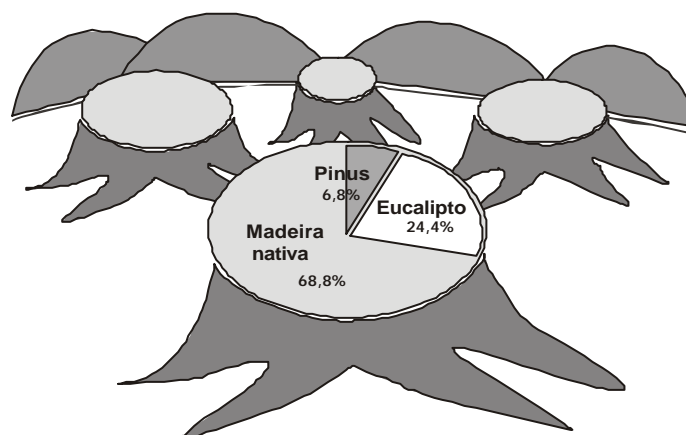
Evolução da frota nacional de carros à álcool de 1979 à 1987



Os métodos mais eficientes para deixar de fumar segundo 30.000 fumantes entrevistados no Canadá



Devastação Selvagem: extração de madeiras no Brasil



### 3.5 Gráfico Polar

É o tipo de gráfico ideal para representar séries temporais cíclicas, ou seja, toda a série que apresenta uma determinada periodicidade.

#### 4.5.1 Como construir um gráfico polar

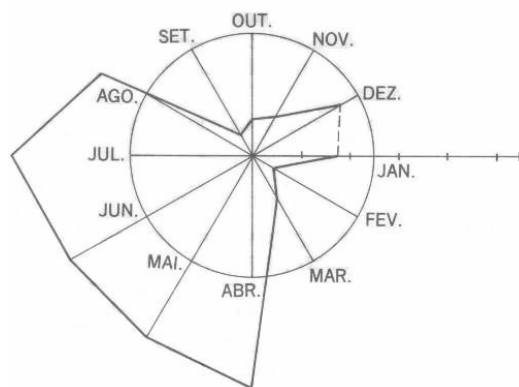
- 1) Traça-se uma circunferência de raio arbitrário (preferencialmente, a um raio de comprimento proporcional a média dos valores da série);
- 2) Constrói-se uma semi-reta (de preferência horizontal) partindo do ponto **0** (pólo) e com uma escala (eixo polar);
- 3) Divide-se a circunferência em tantos arcos forem as unidades temporais;
- 4) Traça-se semi-retas a partir do ponto **0** (pólo) passando pelos pontos de divisão;
- 5) Marca-se os valores correspondentes da variável, iniciando pela semi-reta horizontal (eixo polar);
- 6) Ligam-se os pontos encontrados com segmentos de reta;
- 7) Para fechar o polígono obtido, emprega-se uma linha interrompida.

Precipitação pluviométrica do município de Santa Maria – RS- 1999

Meses	Precipitação (mm)
Janeiro	174,8
Fevereiro	36,9
Março	83,9
Abril	462,7
Maio	418,1
Junho	418,4
Julho	538,7
Agosto	323,8
Setembro	39,7
Outubro	66,1
Novembro	83,3
Dezembro	201,2

Fonte: Base Aérea de Santa Maria

Precipitação pluviométrica do município de Santa Maria – RS- 1999



Fonte: Base Aérea de Santa Maria

Média = 237,31 mm

### 3.6 Cartograma

É a representação de uma carta geográfica. Este tipo de gráfico é empregado quando o objetivo é o de figurar os dados estatísticos diretamente relacionados com as áreas geográficas ou políticas

Dados absolutos (população) – usa-se pontos proporcionais aos dados.

Dados relativos (densidade) – usa-se hachaduras.

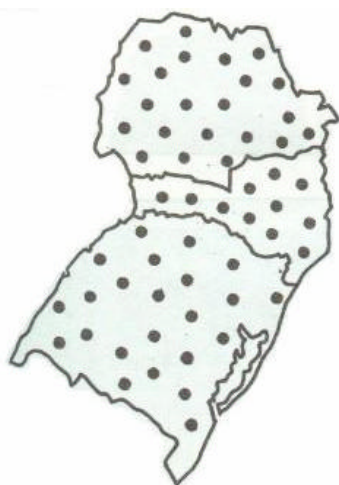
Exemplo:

População da Região Sul do Brasil - 1990

Estado	População (hab.)	Área (m <sup>2</sup> )	Densidade
Paraná	9.137.700	199.324	45,8
Santa Catarina	4.461.400	95.318	46,8
Rio Grande do Sul	9.163.200	280.674	32,6

Fonte: IBGE

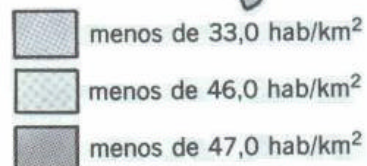
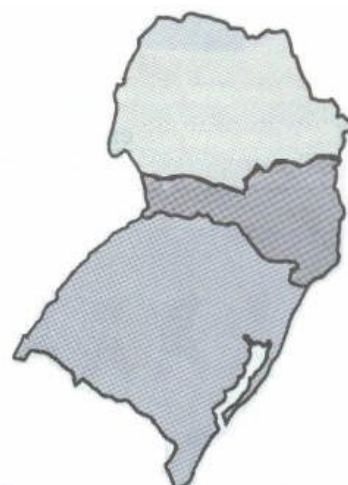
População da Região Sul do Brasil – 1990



● 400.000 habitantes

Fonte: IBGE

Densidade populacional da Região Sul do Brasil – 1990

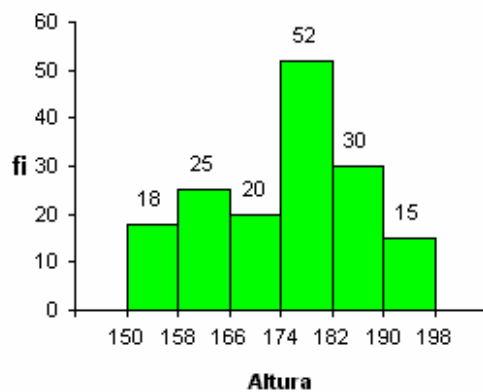


Fonte: IBGE

### 3.7 Gráficos utilizados para a análise de uma distribuição de frequência

#### 3.7.1 Histograma

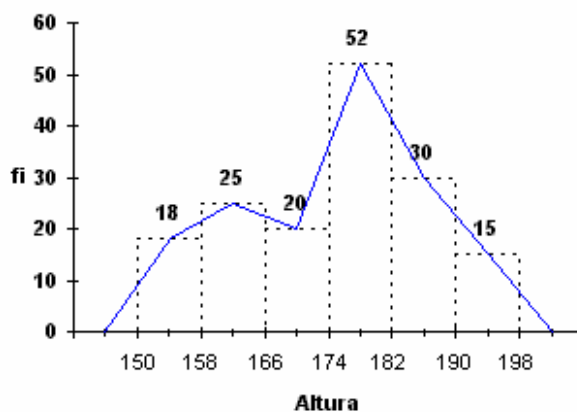
Altura em centímetros de 160 alunos do Curso de Administração da UFSM- 1990



Fonte: Departamento de Estatística (1990)

#### 3.7.2 Polígono de Frequências

Altura em centímetros de 160 alunos do Curso de Administração da UFSM - 1990

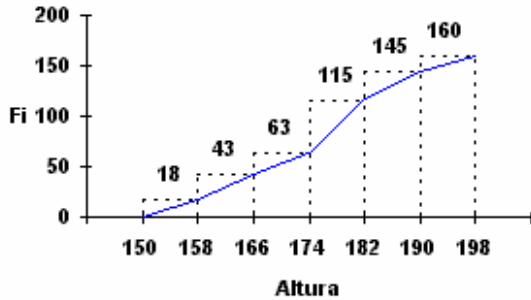


Fonte: Departamento de Estatística (1990)

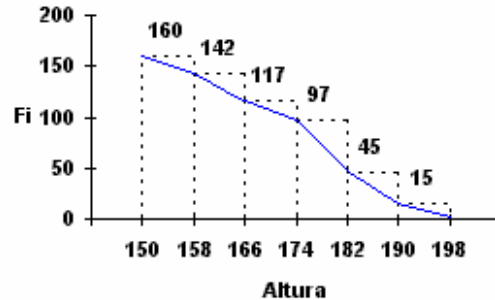
### 3.6.3 Ogivas

Altura em centímetros de 160 alunos do Curso de Administração da UFSM – 1990

Ogiva Crescente



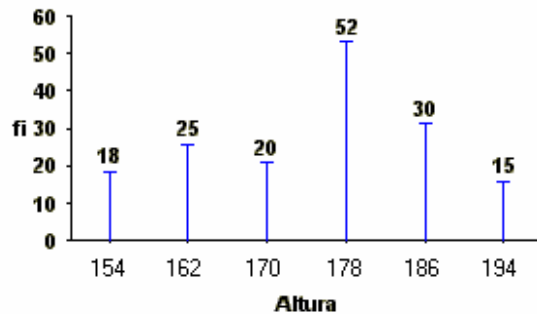
Ogiva Decrescente



### 3.7.4 Gráfico em segmentos de reta vertical

É utilizado para representar uma distribuição de frequência pontual, onde os segmentos de reta são proporcionais às respectivas frequências absolutas.

Altura em centímetros de 160 alunos do Curso de Administração da UFSM- 1990



Fonte: Departamento de Estatística (1990)

### 3.7.5 Como se interpreta um histograma?

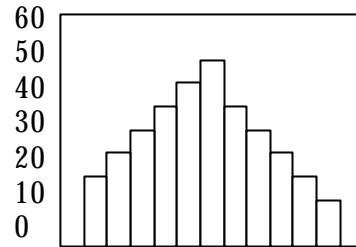
A representação gráfica da distribuição da variável, por **histogramas**. Este gráfico é utilizado para variáveis contínuas.

*Características:*

- Cada barra representa a frequência do intervalo respectivo;
- Os intervalos devem ter a mesma amplitude;
- As barras devem estar todas juntas.

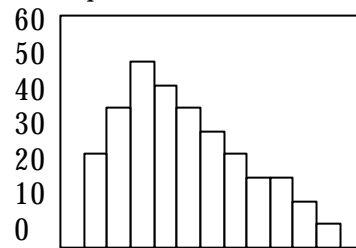
A simples observação da forma do histograma permite algumas conclusões. Veja a figura 4.1. A medida dos dados está no centro do desenho. As freqüências mais altas também estão no centro da figura. Nos processos industriais, esta é a forma desejável.

Figura 4.1 Histograma



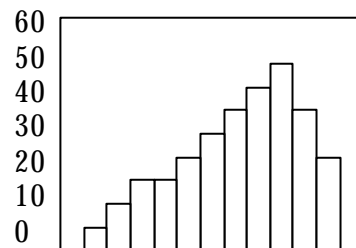
A figura 4.2 apresenta um histograma com assimetria positiva. A média dos dados está localizada à esquerda do centro da figura e a cauda à direita é alongada. Esta ocorre quando o limite inferior é controlado ou quando não podem ocorrer valores abaixo de determinado limite.

Figura 4.2 Histograma com assimetria positiva



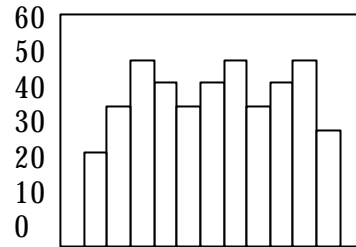
A figura 4.3 apresenta um histograma com assimetria negativa. A média dos dados está localizada à direita do centro da figura e a cauda à esquerda é alongada. Esta forma ocorre quando o limite superior é controlado ou quando não podem ocorrer valores acima de certo limite.

Figura 4.3 Histograma com assimetria negativa



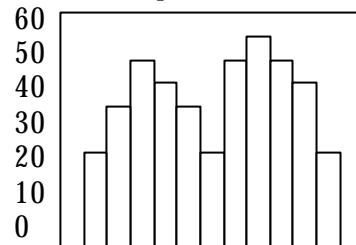
A figura 4.4 mostra um histograma em plateau, Isto é, com exceção das primeiras e das últimas classes, todas as outras têm freqüências quase iguais. Essa forma ocorre quando se misturam várias distribuições com diferentes médias.

Figura 4.4 Histograma em plateau



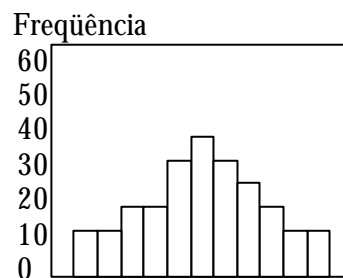
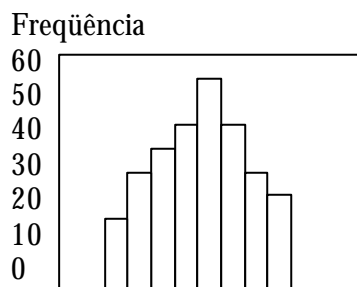
A figura 4.5 mostra um histograma com *dois picos*, ou duas modas. As freqüências são baixas no centro da figura, mas existem dois picos fora do centro. Esta forma ocorre quando duas distribuições com médias bem diferentes se misturam. Podem estar misturados, por exemplo, os produtos de dois turnos de trabalho.

Figura 4.5 Histograma com dois picos



Os histogramas também mostram o grau de dispersão da variável. Veja a figura 4.6. O histograma à esquerda mostra pouca dispersão, mas o histograma à direita mostra grande dispersão.

Figura 4.6 Histogramas com dispersões diferentes



### 3.7.6 Curva de frequência – curva polida

Como, em geral, os dados coletados pertencem a uma amostra extraída de uma população, pode-se imaginar as amostras tornando-se cada vez mais amplas e a amplitude das classes ficando cada vez menor, o que nos permite concluir que o contorno do polígono de frequências tende a se transformar numa curva (curva de frequência), mostrando, de modo mais evidente, a verdadeira natureza da distribuição da população.

Pode-se dizer, então, que, enquanto que o polígono de frequência nos dá a **imagem real** do fenômeno estudado, a curva de frequência nos dá a **imagem tendenciosa**.

Assim, após o traçado de um polígono de frequência, é desejável, muitas vezes, que se faça um **polimento** de modo a mostrar o que seria tal polígono com um número maior de dados.

Esse procedimento é claro, não nos dará certeza absoluta que a **curva polida** seja tal qual a curva resultante de um grande número de dados. Porém, pode-se afirmar que ela assemelha-se mais a curva de frequência que o polígono de frequência obtido de uma amostra limitada.

O polimento, geometricamente, corresponde à eliminação dos vértices da linha poligonal. Consegue-se isso com a seguinte fórmula:

$$fc_i = \frac{f_{\text{ant.}} + 2f_i + f_{\text{post.}}}{4}$$

onde:

$fc_i$  = frequência calculada da classe considerada;

$f_i$  = frequência absoluta da classe  $i$ ;

$f_{\text{ant.}}$  = frequência absoluta da classe anterior a  $i$ ;

$f_{\text{post.}}$  = frequência absoluta da classe posterior a  $i$ ;



**Quando for em fazer o uso da curva polida convém mostrar as frequências absolutas, por meio de um pequeno círculo, de modo que qualquer interessado possa julgar se esse ponto se o ponto é um dado original ou um dado polido.**

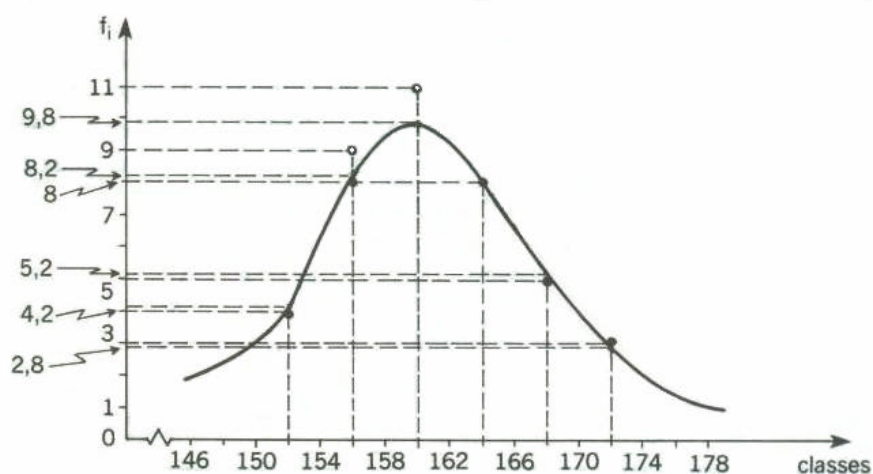
Altura em centímetros de 40 alunas do Curso de Enfermagem da UFSM - 1996

Altura (cm)		$X_i$	$f_i$	$fc_i$
150	--- 154	152	4	$(0 + 2 \times 4 + 9)/4 = 4,25$
154	--- 158	156	9	$(4 + 2 \times 9 + 11)/4 = 8,25$
158	--- 162	160	11	$(9 + 2 \times 11 + 8)/4 = 9,75$
162	--- 166	164	8	$(11 + 2 \times 8 + 5)/4 = 8,00$
166	--- 170	168	5	$(8 + 2 \times 5 + 3)/4 = 5,25$
170	--- 174	172	3	$(5 + 2 \times 3 + 0)/4 = 2,75$
$\Sigma$		----	40	---

Fonte: Departamento de Estatística (1997)



Altura em centímetros de 40 alunas do Curso de Enfermagem da UFSM - 1996



Fonte: Departamento de Estatística (1997)

### 3.7.7 Curvas em forma de sino

As curvas em forma de sino caracterizam-se pelo fato de apresentarem um valor máximo na região central.

Distinguem-se as curvas em forma de sino em: **simétrica** e **assimétrica**

#### a) Curva simétrica

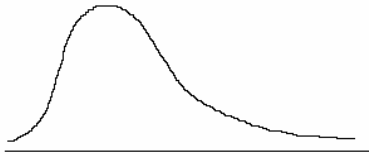
Esta curva caracteriza-se por apresentar o valor máximo no ponto central e os pontos equidistantes desse ponto terem a mesma frequência.



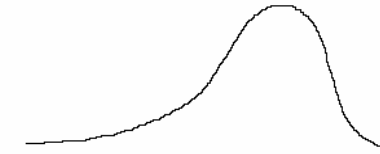
## b) Curvas assimétricas

Na prática, não se encontram distribuições perfeitamente simétricas. As distribuições obtidas de medidas reais são mais ou menos assimétricas, em relação à frequência máxima. Assim, as curvas correspondentes a tais distribuições apresentam a **cauda** de um lado da ordenada máxima mais longa do que o outro. Se a cauda mais longa fica a direita é chamada **assimétrica positiva** ou **enviesada à direita**, se a cauda se alonga a esquerda, a curva é chamada **assimétrica negativa** ou **enviesada à esquerda**.

Assimétrica Positiva



Assimétrica Negativa

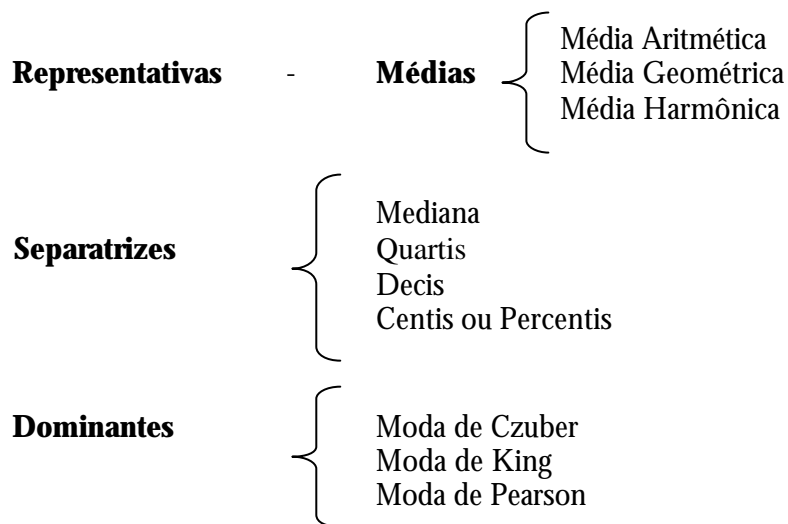


## 4 Medidas Descritivas

Tem por objetivo descrever um conjunto de dados de forma organizada e compacta que possibilita a visualização do conjunto estudado por meio de suas estatísticas, o que não significa que estes cálculos e conclusões possam ser levados para a população.

Podemos classificar as medidas de posição conforme o esquema abaixo:

### 4.1 Medidas de Posição



#### 4.1.1 Representativas (Médias)

São medidas descritivas que tem por finalidade representar um conjunto de dados.

**a) Média Aritmética:** Amostral ( $\bar{X}$ ); Populacional ( $\mu$ )

Dados Não Tabelados

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad \text{ou} \quad m = \frac{\sum_{i=1}^N X_i}{N}$$

Dados Tabelados com Valores Ponderados

Média Aritmética Ponderada ( $X_w$ ), (onde  $W_i$  é o peso)

Nota do aluno "X"  
1º semestre de 1994 - UFSM

Notas ( $X_i$ )	Pesos ( $W_i$ )
7.8	2
8.3	3
8.2	2
5.8	3
$\Sigma$	10

$$X_w = \frac{\sum_{i=1}^n X_i \cdot W_i}{\sum_{i=1}^n W_i}$$

Fonte: Dados Hipotéticos

*Distribuição de frequências*

- **Média Aritmética** ( $\bar{X}$ )

Altura em centímetros de 160 alunos do  
Curso de Administração da UFSM - 1990

Altura (cm)	$X_i$	$f_i$	$X_i \cdot f_i$
150	158	18	2763,0
158	166	25	4037,5
166	174	20	3390,0
174	182	52	9230,0
182	190	30	5565,0
190	198	15	2917,55
$\Sigma$	----	160	27903

$$\bar{X} = \frac{\sum_{i=1}^n X_i \cdot f_i}{\sum_{i=1}^n f_i}$$

Fonte: Departamento de Estatística (1990)

- Características da Média Aritmética Simples

1ª) A Média Aritmética Simples deverá estar entre o menor e o maior valor observado,

$$X_{\min.} \leq \bar{X} \leq X_{\max.}$$

2ª) A soma algébrica dos desvios calculados entre os valores observados e a média aritmética é igual a zero; desvios =  $d = x_i - m$

$$\sum_{i=1}^n d = \sum_{i=1}^n (x_i - m) = \text{zero}$$

3ª) Somando-se ou subtraindo-se todos os valores ( $X_i$ ) da série por uma constante "k" ( $k \neq 0$ ), a nova média aritmética será igual a média original somada ou subtraída por esta constante "k".

$$\begin{array}{ccc} x_i & & y_i = x_i \pm k \\ \Downarrow & & \Downarrow \\ \bar{X} & & \bar{Y} = \bar{X} \pm k \end{array}$$

4ª) Multiplicando-se ou dividindo-se todos os valores ( $X_i$ ) da série por uma constante "k" ( $k \neq 0$ ), a nova média aritmética será igual a média original multiplicada ou dividida por esta constante "k".

$$\begin{array}{ccc} x_i & y_i = x_i \times k & y_i = x_i \div k \\ \Downarrow & \Downarrow & \Downarrow \\ \bar{X} & \bar{Y} = \bar{X} \times k & \bar{Y} = \bar{X} \div k \end{array}$$

### b) Média Geométrica: ( $X_g$ ):

A aplicação da média geométrica deve ser feita, quando os valores do conjunto de dados considerado se comportam segundo uma progressão geométrica (P.G.) ou dela se aproximam.

#### Dados Não Tabelados

$$X_g = \sqrt[n]{\prod_{i=1}^n X_i} = \sqrt{X_1 \cdot X_2 \cdot \dots \cdot X_n}$$

#### Dados Tabelados

$$X_g = \sqrt[n]{\prod_{i=1}^n X_i^{f_i}} = \sqrt{X_1^{f_1} \cdot X_2^{f_2} \cdot \dots \cdot X_n^{f_n}}$$

Usando um artifício matemático, pode-se usar para calcular a média geométrica a seguinte fórmula:

$$X_g = 10^{\frac{1}{\sum_{i=1}^n f_i} (f_1 \cdot \log X_1 + f_2 \cdot \log X_2 + \dots + f_n \cdot \log X_n)} = 10^{\frac{1}{\sum_{i=1}^n f_i} \sum_{i=1}^n f_i \cdot \log X_i}$$

### c) Média Harmônica ( $X_h$ )

É usada para dados inversamente proporcionais.

Ex.: Velocidade Média, Preço de Custo Médio

#### 4.1.2 Emprego da média

- 1) Deseja-se obter a medida de posição que possui a maior estabilidade;
- 2) Houver necessidade de um tratamento algébrico ulterior.

#### Dados Não Tabelados

$$X_h = \frac{n}{\sum_{i=1}^n \frac{1}{X_i}} = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_n}}$$

#### Dados Tabelados

$$X_h = \frac{\sum_{i=1}^n f_i}{\sum_{i=1}^n \frac{f_i}{X_i}} = \frac{f_1 + f_2 + \dots + f_n}{\frac{f_1}{X_1} + \frac{f_2}{X_2} + \dots + \frac{f_n}{X_n}}$$



Deve-se observar esta propriedade entre as médias  $\bar{X} \geq X_g \geq X_h$

#### 4.1.3 Separatrizes (Mediana, Quartis, Decis e Centis ou Percentis)

São medidas de posição que divide o conjunto de dados em partes proporcionais, quando os mesmos são ordenados.

##### a) Dados não tabelados

Antes de determinarmos as separatrizes devemos em primeiro lugar encontrar a posição da mesma.

- Se o número de elementos for par ou ímpar, as separatrizes seguem a seguinte ordem:

$$\text{Posição} = \frac{i(n+1)}{S}$$

$$\text{se for mediana} \begin{cases} i=1 \\ S=2 \end{cases}$$

$$\text{se for quartis} \begin{cases} 1 \leq i \leq 3 \\ S=4 \end{cases}$$

$$\text{se for } \mathbf{decis} \begin{cases} 1 \leq i \leq 9 \\ S = 10 \end{cases}$$

$$\text{se for } \mathbf{centis} \begin{cases} 1 \leq i \leq 99 \\ S = 100 \end{cases}$$

### Dados Tabelados

**b) Distribuição de freqüências pontual: segue a mesma regra usada para dados não tabelados**

**c) Distribuição de freqüências intervalar**

$$S_i = l_{S_i} + \frac{\left( \frac{i.n}{S} - F_{\text{ant}} \right) \cdot h}{f_{S_i}}$$

onde:

$$S_i = Md \Rightarrow i = 1;$$

$$S_i = Q_i \Rightarrow 1 \leq i \leq 3;$$

$$S_i = D_i \Rightarrow 1 \leq i \leq 9;$$

$$S_i = C_i \text{ ou } P_i \Rightarrow 1 \leq i \leq 99$$

$$l_{S_i} \Rightarrow \text{limite inferior da classe que contém a separatriz;}$$

$$\frac{i.n}{S} \Rightarrow \text{posição da separatriz;}$$

$$F_{\text{ant}} \Rightarrow \text{freqüência acumulada da classe anterior a que contém a separatriz;}$$

$$h \Rightarrow \text{amplitude do intervalo de classe;}$$

$$f_{S_i} \Rightarrow \text{freqüência absoluta da classe que contém a separatriz;}$$

#### **4.1.4 Emprego da mediana**

- 1) Quando se deseja obter um ponto que divide a distribuição em partes iguais;
- 2) Há valores extremos que afetam de uma maneira acentuada a média;
- 3) A variável em estudo é salário.

#### **4.1.5 Dominantes - Moda (Mo)**

É definida como sendo a observação de maior freqüência.

### a) Dados não tabelados

Ex.: 3 4 4 4 5 5 6 6 7 8 9	$\Rightarrow Mo = 4$ (unimodal)
5 6 7 8 9 10 11 12 13	$\Rightarrow Mo = \bar{x}$ (amodal)
1 1 2 2 3 3 3 4 5 5 5	$\Rightarrow Mo_1 = 3 \ Mo_2 = 5$ (bimodal)
5 5 6 6 7 7 8 8	$\Rightarrow Mo = \bar{x}$ (amodal)
5 5 6 6 7 7 8	$\Rightarrow Mo_1 = 5 \ Mo_2 = 6 \ Mo_3 = 7$ (multimodal)



Acima de 3 modas usamos o termo multimodal.

### Dados Tabelados

#### a) Distribuição de freqüências pontual

- **Moda Bruta** ( $Mo_b$ ): é o ponto médio da classe de maior freqüência

$$Mo_b = X_i$$

#### b) Distribuição de freqüências intervalar

- **Moda de Czuber** ( $Mo_c$ ): O processo para determinar a moda usado por Czuber leva em consideração as freqüências anteriores e posteriores à classe modal.

$$Mo_c = l_{Mo} + \left( \frac{\Delta_1}{\Delta_1 + \Delta_2} \right) h \Rightarrow \begin{cases} \Delta_1 = f_{Mo} - f_{ant} \\ \Delta_2 = f_{Mo} - f_{pos} \end{cases}$$

onde:

$l_{Mo}$   $\Rightarrow$  limite inferior da classe modal;

$f_{Mo}$   $\Rightarrow$  freqüência absoluta da classe modal;

$h$   $\Rightarrow$  amplitude do intervalo de classe;

$f_{ant}$   $\Rightarrow$  freqüência absoluta da classe anterior a classe modal;

$f_{pos}$   $\Rightarrow$  freqüência absoluta da classe posterior a classe modal;


- **Moda de King** ( $Mo_k$ ): O processo proposto por King considera a influência existente das classes anterior e posterior sobre a classe modal. A inconveniência deste processo é justamente não levar em consideração a freqüência máxima.

$$Mo_k = l_{Mo} + \left( \frac{f_{pos}}{f_{pos} + f_{ant}} \right) h$$



- **Moda de Pearson** ( $Mo_p$ ): O processo usado por Pearson pressupõe que a distribuição seja aproximadamente simétrica, na qual a média aritmética e a mediana são levadas em consideração.

$$Mo_p = 3 Md - 2 \bar{X}$$

 Um distribuição é considerada simétrica quando  $\bar{X} \equiv Md \equiv Mo$ .

#### 4.1.6 Emprego da moda

- 1) Quando se deseja obter uma medida rápida e aproximada de posição;
- 2) Quando a medida de posição deve ser o valor mais típico da distribuição.

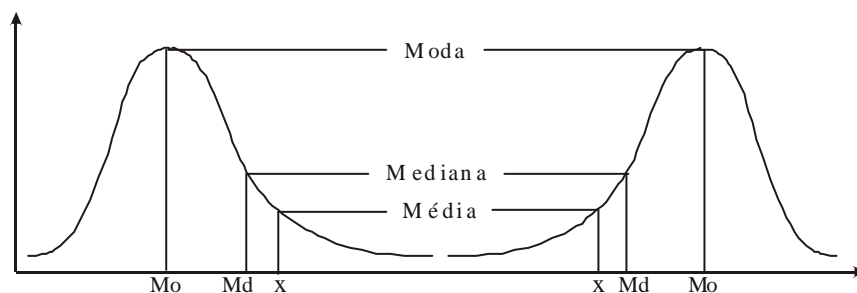
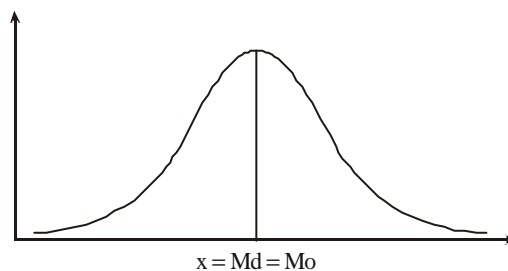
#### 4.1.7 Posição relativa da média, mediana e moda

Quando uma distribuição é simétrica, as três medidas coincidem. Porém, a assimetria torna-as diferentes e essa diferença é tanto maior quanto maior é a assimetria. Assim, em uma distribuição temos:

$\bar{x} = Md = Mo \rightarrow$  curva simétrica

$\bar{x} < Md < Mo \rightarrow$  curva assimétrica negativa

$Mo < Md < \bar{x} \rightarrow$  curva assimétrica positiva



### 4.1.8 Exercícios

Para os dados abaixo calcule: Md; Q<sub>1</sub>; Q<sub>3</sub>; D<sub>3</sub>; C<sub>70</sub>

1)

3    7    8    10    12    13    15    17    18    21  
4    6    8    11    13    14    17    18    19    22    25

2)

Alturas dos alunos da Turma "X"  
no 1<sup>a</sup> sem. de 1994 - UFSM

Alturas	f <sub>i</sub>
63	15
75	25
84	30
91	20
Σ	90

Fonte: Dados Hipotéticos

3)

Alturas dos alunos da Turma "Y"  
no 1<sup>a</sup> sem. de 1994 - UFSM

Alturas	f <sub>i</sub>	F <sub>i</sub>
61	12	12
65	23	35
69	34	69
73	26	95
77	15	110
Σ	110	110

Fonte: Dados Hipotéticos

### 4.2 Medidas de Variabilidade ou Dispersão

Visam descrever os dados no sentido de informar o grau de dispersão ou afastamento dos valores observados em torno de um valor central representativo chamado média. Informa se um conjunto de dados é homogêneo (pouca variabilidade) ou heterogêneo (muita variabilidade).

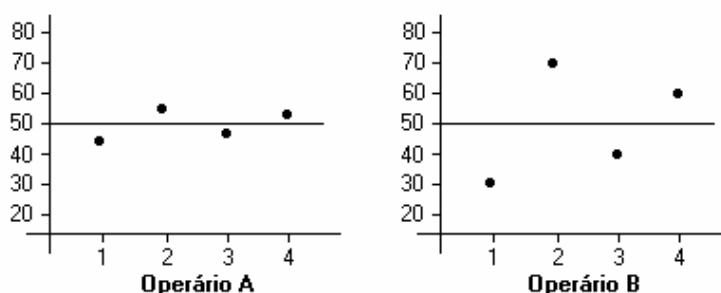
As medidas de dispersão podem ser:

**Absoluta** {  
- Desvio extremo - amplitude  
- Desvio Médio  
- Desvio Padrão  
- Desvio quadrático - Variância

Para estudarmos as medidas de variabilidade para dados não tabelados usaremos um exemplo prático. Supomos que uma empresa esteja querendo contratar um funcionário, e no final da concorrência sobraram dois candidatos para uma única vaga. Então foi dado 4 tarefas para cada um, onde as mesmas tiveram como registro o tempo (em minutos) de execução.

TAREFAS	1	2	3	4
OPERARIO 1 (TEMPO)	55	45	52	48
OPERARIO 2 (TEMPO)	30	70	40	60

### - Análise Gráfica



### - Medidas de dispersão Absoluta:

- **Desvio Extremo ou Amplitude de Variação (H):** É a diferença entre o maior e o menor valor de um conjunto de dados

$$H = X_{\max} - X_{\min}$$

- **Desvio Médio ( $\bar{d}$ ):** Em virtude do  $\sum_{i=1}^n (X_i - \bar{X}) = 0$ , usamos para calcular o desvio

médio  $\sum_{i=1}^n |X_i - \bar{X}| = 0$ , assim ficando:

Para dados não tabelados

$$\bar{d} = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n} = \frac{|X_1 - \bar{X}| + |X_2 - \bar{X}| + \dots + |X_n - \bar{X}|}{n}$$

Para dados tabelados

$$\bar{d} = \frac{\sum_{i=1}^n (f_i |X_i - \bar{X}|)}{\sum_{i=1}^n f_i} = \frac{f_1 |X_1 - \bar{X}| + f_2 |X_2 - \bar{X}| + \dots + f_n |X_n - \bar{X}|}{\sum_{i=1}^n f_i}$$

- **Desvio Quadrático ou Variância** :  $S^2$  (amostra) ou  $\sigma^2$  (população)

Para dados não tabelados:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n}$$

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n-1}$$

Para dados tabelados

$$s^2 = \frac{\sum_{i=1}^n f_i (X_i - \bar{X})^2}{\sum_{i=1}^n f_i} = \frac{f_1 (X_1 - \bar{X})^2 + f_2 (X_2 - \bar{X})^2 + \dots + f_n (X_n - \bar{X})^2}{\sum_{i=1}^n f_i}$$

$$S^2 = \frac{\sum_{i=1}^n f_i (X_i - \bar{X})^2}{\sum_{i=1}^n f_i - 1} = \frac{f_1 (X_1 - \bar{X})^2 + f_2 (X_2 - \bar{X})^2 + \dots + f_n (X_n - \bar{X})^2}{\sum_{i=1}^n f_i - 1}$$

- **Desvio Padrão:**  $s$  (amostra) ou  $\sigma$  (população)

Para dados não tabelados:


$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}} = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n}}$$

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n-1}}$$

Para dados tabelados

$$s = \sqrt{\frac{\sum_{i=1}^n f_i (X_i - \bar{X})^2}{\sum_{i=1}^n f_i}} = \sqrt{\frac{f_1 (X_1 - \bar{X})^2 + f_2 (X_2 - \bar{X})^2 + \dots + f_n (X_n - \bar{X})^2}{\sum_{i=1}^n f_i}}$$

$$S = \sqrt{\frac{\sum_{i=1}^n f_i (X_i - \bar{X})^2}{\sum_{i=1}^n f_i - 1}} = \sqrt{\frac{f_1 (X_1 - \bar{X})^2 + f_2 (X_2 - \bar{X})^2 + \dots + f_n (X_n - \bar{X})^2}{\sum_{i=1}^n f_i - 1}}$$

  $(n - 1)$  é usado como um fator de correção, onde devemos considerar a variância amostral como uma estimativa da variância populacional.

- Propriedades da Variância

- 1ª) Somando-se ou subtraindo-se uma constante  $k$  a cada valor observado a variância não será alterada;
- 2ª) Multiplicando-se ou dividindo-se por uma constante  $k$  cada valor observado a variância ficará multiplicada ou dividida pelo quadrado dessa constante.

**Outra forma de calcular o desvio padrão**

O desvio padrão mede bem a dispersão de um conjunto de dados, mas é difícil de calcular. Então, você pode obter o desvio padrão através da seguinte relação:

$$\hat{\sigma} = \frac{R}{d_2}$$

onde  $R$  é a amplitude e o valor de  $d_2$ , que depende do tamanho da amostra, é encontrado na tabela a seguir. Este método de calcular o desvio padrão fornece boas estimativas para amostras de pequeno tamanho ( $n=4, 5$  ou  $6$ ), mas perde a eficiência se  $n > 10$ . De qualquer

forma, é essa relação entre a amplitude e o desvio padrão de uma amostra que permite fazer gráficos de controle  $\bar{X} - R$ .

TABELA 1: - Fatores para construir um gráfico de controle

n	2	3	4	5	6	7	8	9	10	11	12	13	14
d <sub>2</sub>	1,128	1,693	2,059	2,326	2,534	2,704	2,847	2,970	3,078	3,173	3,258	3,336	3,407
n	15	16	17	18	19	20	21	22	23	24	25	---	---
d <sub>2</sub>	0,347	3,532	3,532	3,640	3,689	3,735	3,778	3,819	3,858	3,391	3,391	---	---

Fonte: Montgomery, D.C. Statical Quality Control. Nova York, Wyley. 1991.

### 4.3 Medidas de Dispersão Relativa

$$\mathbf{Relativa} \left\{ \begin{array}{l} - \text{Variância relativa} \\ - \text{Coeficiente de Variação (Pearson)} \end{array} \right.$$


É a medida de variabilidade que em geral é expressa em porcentagem, e tem por função determinar o grau de concentração dos dados em torno da média, geralmente utilizada para se fazer a comparação entre dois conjuntos de dados em termos percentuais, esta comparação revelará o quanto os dados estão próximos ou distantes da média do conjunto de dados.

#### - Variância Relativa

$$V.R. = \frac{\sigma^2}{\mu^2} \text{ ou } \frac{S^2}{\bar{X}^2}$$

#### - Coeficiente de Variação de Pearson

$$C.V. = \frac{\sigma}{\mu} \text{ ou } \frac{S}{\bar{X}} \times 100$$

 C.V. ≤ 50% ⇒ a média é representativa

C.V. ≅ 0 ⇒ é a maior representatividade da média (S = 0)

### 4.4 Momentos, assimetria e curtose

As medidas de assimetria e curtose complementam as medidas de posição e de dispersão no sentido de proporcionar uma descrição e compreensão mais completa das distribuições de freqüências. Estas distribuições não diferem apenas quanto ao valor médio e à variabilidade, mas também quanto a sua forma (assimetria e curtose).

Para estudar as medidas de assimetria e curtose, é necessário o conhecimento de certas quantidades, conhecidas como momentos.

#### 4.4.1 Momentos

São medidas descritivas de caráter mais geral e dão origem às demais medidas descritivas, como as de tendência central, dispersão, assimetria e curtose. Conforme a potência considerada tem-se a ordem ou o grau do momento calculado.

##### - Momentos simples ou centrados na origem ( $m_r$ )

O momento simples de ordem “r” é definido como:

$$m_r = \frac{\sum X_i^r}{n}, \quad \text{para dados não tabelados;}$$

$$m_r = \frac{\sum X_i^r f_i}{\sum f_i}, \quad \text{para dados tabelados.}$$

onde:

r é um número inteiro positivo;

$m_0 = 1$ ;

$m_1 =$  média aritmética.

##### - Momentos centrados na média ( $M_r$ )

O momento de ordem “r” centrado na média, é definido como:

$$M_r = \frac{\sum (X_i - \bar{X})^r}{n} = \frac{\sum d_i^r}{n}, \quad \text{para dados não-tabelados}$$

$$M_r = \frac{\sum (X_i - \bar{X})^r f_i}{\sum f_i} = \frac{\sum d_i^r f_i}{n}, \quad \text{para dados tabelados}$$

onde:  $M_0 = 1$ ;

$M_1 = 0$ ;

$M_2 =$  variância ( $s^2$ ).

#### - Momentos abstratos ( $\alpha_r$ )

São definidos da seguinte forma:

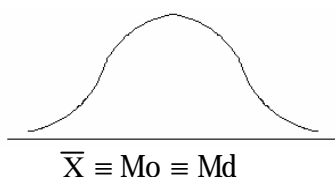
$$\alpha_r = \frac{M_r}{s_r}$$

onde:  $s$  = desvio padrão.

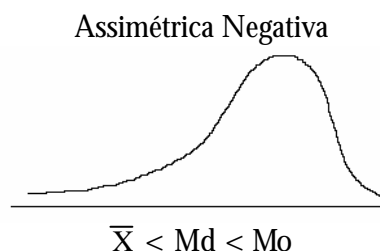
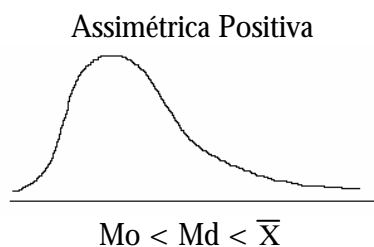
#### 4.4.2 Assimetria

Uma distribuição de valores sempre poderá ser representada por uma curva (gráfico). Essa curva, conforme a distribuição, pode apresentar várias formas. Se considerarmos o valor da moda da distribuição como ponto de referência, vemos que esse ponto sempre corresponde ao valor de ordenada máxima, dando-nos o ponto mais alto da curva representativa da distribuição considerada, logo a curva será analisada quanto à sua assimetria.

- **Distribuição Simétrica:** É aquela que apresenta a  $\bar{X} \equiv Mo \equiv Md$  e os quartis  $Q_1$  e  $Q_3$  equidistantes do  $Q_2$ .



#### - Distribuição Assimétrica



Podemos medir a assimetria de uma distribuição, calculando os coeficientes de assimetria. Sendo o mais utilizado o Coeficiente de Assimetria de Pearson.

$$As = \frac{\bar{X} - Mo}{S}$$

- Se  $As < 0 \Rightarrow$  a distribuição será Assimétrica Negativa;
- Se  $As > 0 \Rightarrow$  a distribuição será Assimétrica Positiva;
- Se  $As = 0 \Rightarrow$  a distribuição será Simétrica.



Quando não tivermos condições de calcularmos o desvio padrão podemos usar a seguinte fórmula:

$$A_s = \frac{Q_3 + Q_1 - 2Md}{Q_3 - Q_1}$$

- **Coefficiente momento de assimetria** ( $\alpha_3$ ): É o terceiro momento abstrato.

$$\alpha_3 = \frac{M_3}{s^3}$$

O campo de variação do coeficiente de assimetria é:  $-1 \leq \alpha_3 \leq +1$

- **Intensidade da assimetria:**

- |                          |               |                   |
|--------------------------|---------------|-------------------|
| $ \alpha_3  < 0,2$       | $\Rightarrow$ | simetria;         |
| $0,2 <  \alpha_3  < 1,0$ | $\Rightarrow$ | assimetria fraca; |
| $ \alpha_3  > 1,0$       | $\Rightarrow$ | assimetria forte. |

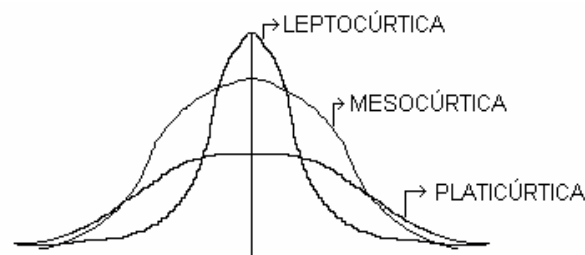
#### 4.4.3 Curtose

Já apreciamos as medidas de tendência central, de dispersão e de assimetria. Falta somente examinarmos mais uma das medidas de uso comum em Estatística, para se positivarem as características de uma distribuição de valores: são as chamadas Medidas de Curtose ou de Achatamento, que nos mostra até que ponto a curva representativa de uma distribuição é a mais aguda ou a mais achatada do que uma curva normal, de altura média.

- **Curva Mesocúrtica (Normal):** É considerada a curva padrão.

- **Curva Leptocúrtica:** É uma curva mais alta do que a normal. Apresenta o topo relativamente alto, significando que os valores se acham mais agrupados em torno da moda.

- **Curva Platicúrtica:** É uma curva mais baixa do que a normal. Apresenta o topo achatado, significando que várias classes apresentam frequências quase iguais.



**- Coeficiente de Curtose**

$$K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$

- Se  $K > 0.263 \Rightarrow$  a distribuição será Platicúrtica.
- Se  $K = 0.263 \Rightarrow$  a distribuição será Mesocúrtica;
- Se  $K < 0.263 \Rightarrow$  a distribuição será Leptocúrtica;

**Coefficiente momento de curtose ( $\alpha_4$ ):** Corresponde ao momento abstrato de quarta ordem.

$$\alpha_4 = \frac{M_4}{s^4}$$

onde:  $M_4$  = momento centrado de quarta ordem.

**Interpretação:**

- Se  $\alpha_4 < 3 \Rightarrow$  curva Platicúrtica;
- Se  $\alpha_4 = 3 \Rightarrow$  curva Mesocúrtica;
- Se  $\alpha_4 > 3 \Rightarrow$  curva Leptocúrtica.

**4.5 Exercícios**

Para os exercícios abaixo construa uma tabela de dispersão o suficiente para determinar as medidas de posição (média aritmética, mediana e moda de czuber), dispersão (desvio padrão e variância, coeficiente de variação de Pearson), assimetria (coeficiente de assimetria, e coeficiente de curtose). Faça um relatório referente ao comportamento dos dados em função dos resultados obtidos.

1) De um exame final de Estatística, aplicado a 50 alunos da Universidade Luterana, Ano 1999 resultaram as seguintes notas:

4,0	4,2	4,3	4,4	4,5	4,5	4,6	5,0	5,1	5,2
5,3	5,3	5,5	5,7	5,8	6,0	6,1	6,3	6,4	6,5
6,6	6,7	6,8	6,9	7,0	7,2	7,5	7,6	7,7	7,9
8,0	8,3	8,5	8,6	8,8	8,9	9,0	9,1	9,2	9,3
9,3	9,4	9,4	9,5	9,5	9,6	9,7	9,8	9,9	10,0

2) Os dados a seguir refere-se a altura em centímetros de 70 alunos da PUCC, turma 6, ano 2000.

153	154	155	156	158	160	160	161	161	161
162	162	163	163	164	164	165	166	167	167
168	168	169	169	170	170	170	171	171	172
172	173	173	174	174	175	175	176	177	178
179	179	180	182	183	184	185	186	186	187
188	188	189	189	190	191	192	192	192	192
193	194	194	195	197	197	199	200	201	205

3) Os dados a seguir referem-se aos salários anuais pagos em dólares a 60 funcionários da Empresa "PETA S.A." em 1997.

50,00	52,50	53,50	54,00	54,20	55,50	56,30	56,50	57,00	58,10
58,50	59,00	60,30	61,50	62,00	62,90	63,50	64,00	64,30	65,00
66,00	66,25	67,50	68,00	68,70	69,50	70,00	72,00	75,00	76,50
77,00	78,00	80,00	81,50	82,50	83,50	85,00	87,30	88,00	89,10
90,00	91,35	92,10	93,20	94,00	95,25	96,00	97,00	98,00	99,80
100,10	100,20	101,00	102,00	103,40	104,30	105,00	107,00	108,00	109,10

# 5 Probabilidade e Variáveis Aleatórias

## 5.1 Modelos Matemáticos

Podem-se distinguir dois tipos de modelos matemáticos:

### 6.1.1 Modelos Determinísticos

Refere-se a um modelo que estipule que as condições sob as quais um experimento seja executado *determinem* o resultado do experimento. O **modelo determinístico** requer o uso de parâmetros pré-determinados em equações que definem processos precisos.

Em outras palavras, um modelo determinístico emprega "Considerações Físicas" para prever resultados.

### 6.1.2 Modelos Não Determinísticos ou Probabilísticos

São aqueles que informam com que chance ou probabilidade os acontecimentos podem ocorrer. Determina o "grau de credibilidade" dos acontecimentos. (Modelos Estocásticos).

Em outras palavras, um modelo probabilístico emprega uma mesma espécie de considerações para especificar uma distribuição de probabilidade.

## 5.2 Conceitos em Probabilidade

Os conceitos fundamentais em probabilidade são experimentos aleatórios, espaço amostral e eventos.

### 5.2.1 Experimento aleatório (W)

Qualquer processo aleatório, capaz de produzir observações, os resultados surgem ao acaso, podendo admitir repetições no futuro. Um experimento aleatório apresenta as seguintes características:

a - os resultados podem repetir-se  $n$  vezes ( $n \rightarrow \infty$ );

b - embora não se possa prever que resultados ocorrerão, pode-se descrever o conjunto de resultados possíveis;

c - a medida que se aumenta o número de repetições, aparece uma certa regularidade nos resultados.

### **5.2.2 Espaço Amostral (S)**

É o conjunto de resultados possíveis, de um experimento aleatório. Quanto ao número de elementos pode ser:

#### **5.2.2.1 Finito**

Número limitado de elementos;

Ex.:  $S = \{1, 2, 3, 4, 5, 6\}$

#### **5.2.2.2 Infinito**

Número ilimitado de elementos, pode ser sub-dividido em:

##### **a - Enumerável**

Quando os possíveis resultados puderem ser postos em concordância biunívoca com o conjunto dos números naturais (N) (caso das variáveis aleatórias discretas).

Ex.: N

##### **b - Não Enumerável**

Quando os possíveis resultados não puderem ser postos em concordância biunívoca com o conjunto dos números naturais (caso das variáveis aleatórias contínuas).

Ex.: R

### **5.2.3 Evento (E)**

Um evento (E) é qualquer subconjunto de um espaço amostral (S).

Pode-se ter operações entre eventos da mesma forma que com conjuntos, como mostra a seguir.

### **5.2.4 Operações com Eventos**

#### **5.2.4.1 A união B**

Símbolo utilizado "U", é o evento que ocorrerá se, e somente se, A ou B ou ambos ocorrerem;

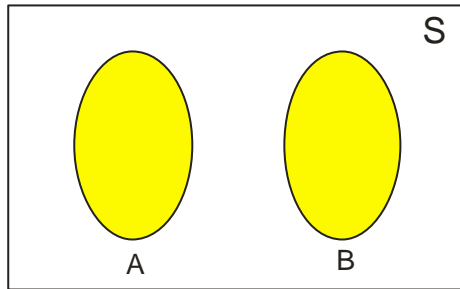


FIGURA 6.1 - Evento A união B

### 5.2.4.2 A interseção B

Símbolo utilizado " $\cap$ ", é o evento que ocorrerá se, e somente se, A e B ocorrem simultaneamente.

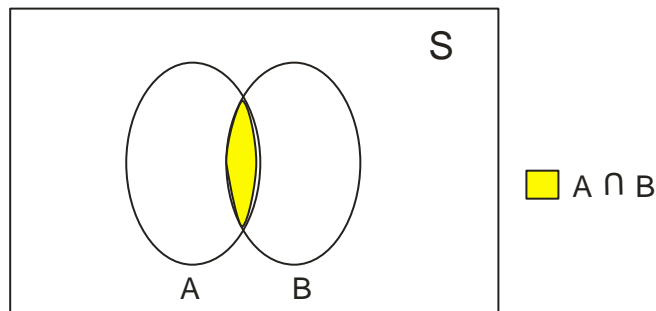


FIGURA 6.2 - Evento A interseção B

### 5.2.4.3 Complementar de A

Simbologia " $\bar{A}$ ", é o evento que ocorrerá se, e somente se A não ocorrer.

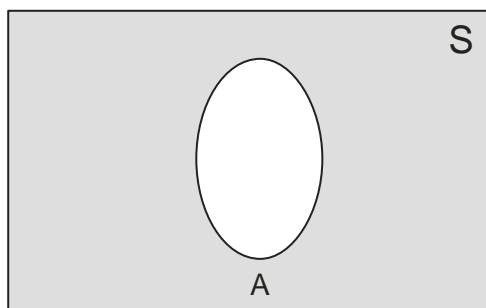


FIGURA 6.3 - Evento complementar de A ( $\bar{A}$ )

## 5.2.5 Tipos de eventos

### 5.2.5.1 Eventos Mutuamente Excludentes

São ditos eventos mutuamente excludentes, quando a ocorrência de um implica ou não ocorrência de outro, isto é, não pode ocorrer juntos, e conseqüentemente,  $A \cap B$  é o conjunto vazio ( $\emptyset$ ).

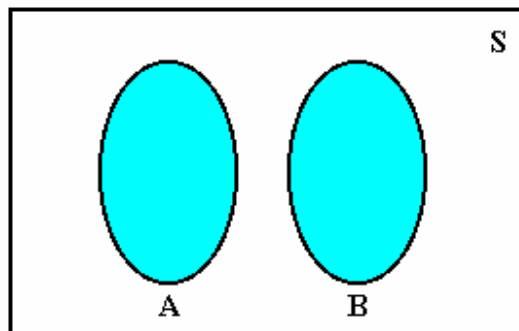


FIGURA 6.4 - Eventos mutuamente excludentes

### 5.2.5.2 Eventos Não Excludentes ou Quaisquer

São ditos eventos não excludentes quando a ocorrência de um implica na ocorrência do outro, isto é, são aqueles que ocorrem ao mesmo tempo,  $A \cap B \neq \emptyset$ .

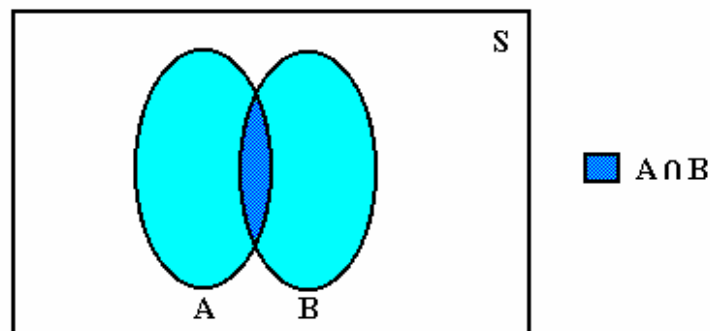


FIGURA 6.5 - Evento não excludentes

### 5.2.5.3 Eventos Independentes

São aqueles cuja ocorrência de um evento, não possui efeito algum na probabilidade de ocorrência do outro.

$A \cap B \neq \emptyset$ , se A e B forem Quaisquer;

$A \cap B = \emptyset$ , se A e B forem Mutuamente Excludentes.

logo,

$$P(A \cap B) = P(A) \cdot P(B)$$

Ex.: A e B eventos Quaisquer

$$S = \{ 1, 2, 3, 4 \}$$

$$A = \{ 1, 2 \}$$

$$B = \{ 2, 4 \}$$

$$A \cap B = \{ 2 \}$$

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(A) = \frac{2}{4}$$

$$P(B) = \frac{2}{4}$$

$$P(A \cap B) = \frac{1}{4}$$

#### 5.2.5.4 Eventos Dependentes ou Condicionados

Existem varias situações onde a ocorrência de um evento pode influenciar fortemente na ocorrência de outro.

Assim, se (A) e (B) são eventos, deseja-se definir uma quantidade denominada probabilidade condicional do evento (A) dado que o evento (B) ocorre, ou sob a forma simbólica  $P\left(\frac{A}{B}\right)$ .

Assim, dá-se a seguinte definição:

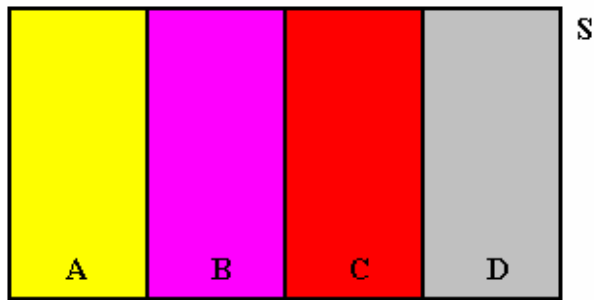
$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$$

onde  $P(B) > 0$ . Se  $P(B) = 0$ , tem-se que  $P\left(\frac{A}{B}\right)$  não é definida.

#### 5.2.5.5 Eventos Coletivamente Exaustivos

São aqueles que ocorrem se nenhum outro ocorrer.





$$A \cap B \cap C \cap D = \emptyset$$

FIGURA 6.6 - Evento coletivamente exaustivos

### 5.3 Conceitos de Probabilidade

#### 5.3.1 Conceito Empírico de Probabilidade

O problema fundamental da probabilidade consiste em atribuir um número a cada evento (E), o qual avaliará quão possível será a ocorrência de "E", quando o experimento for realizado.

Uma possível maneira de tratar a questão seria determinar a frequência relativa do evento E ( $f_r(E)$ ),

$$f_r(E) = \frac{\text{número de ocorrências do evento (E)}}{\text{número de repetições do experimento } (\Omega)}$$

Surgem, no entanto, dois problemas:

- a - Qual deve ser o número de repetições do experimento ( $\Omega$ );
- b - A sorte ou habilidade do experimentador poderá influir nos resultados, de forma tal que a probabilidade é definida como sendo:

$$P(E) = \lim_{n \rightarrow \infty} f_r(E),$$

onde "n" é o número de repetições do experimento  $\Omega$ .

#### 5.3.2 Definição Clássica ou Enfoque "A priori" de Probabilidade

Se existe "a" resultados possíveis favoráveis a ocorrência de um evento "E" e "b" resultados possíveis não favoráveis, sendo os mesmos mutuamente excludentes, então:

$$P(E) = \frac{a}{a+b},$$

onde os resultados devem ser verossímeis (possível e verdadeiro) e permite a observação dos valores da probabilidade antes de ser observado qualquer amostra do evento (E).

### 5.3.3 Definição Axiomática

Seja ( $\Omega$ ) um experimento, seja (S) um espaço amostral associado a ( $\Omega$ ). A cada evento (E) associa-se um número real representado por  $P(E)$  e denominaremos de probabilidade de E, satisfazendo as seguintes propriedades:

a -  $0 \leq P(E) \leq 1$ ;

b -  $P(S) = 1$ ;

c - Se A e B são eventos mutuamente excludentes, então:

$$P(A \cup B) = P(A) + P(B).$$

d - Se  $A_1, A_2, \dots, A_n$  são eventos mutuamente excludentes dois a dois, então:

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$

ou

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i).$$

As propriedades anteriores são conhecidas como axiomas da teoria da probabilidade. Os axiomas, muitas vezes, se inspiram em resultados experimentais e que, assim, definem a probabilidade de forma que possa ser confirmada experimentalmente.

### 5.3.4 Teoremas Fundamentais

Teorema 1 - Se  $\emptyset$  for evento vazio, então  $P(\emptyset) = 0$ .

Prova: Seja um evento  $A = \emptyset$ . Assim,  $A = A \cup \emptyset$ , como  $A \cap \emptyset = \emptyset$ , de acordo com o item (3.2.3.4), A e  $\emptyset$  são mutuamente excludentes, então:

$$P(A) = P(A \cup \emptyset)$$

$$P(A) = P(A) + P(\emptyset)$$

$$P(\emptyset) = P(A) - P(A)$$

$$P(\emptyset) = 0.$$

Teorema 2 - Se o evento  $\bar{A}$  for o evento complementar de A, então  $P(\bar{A})=1-P(A)$ .

Prova:  $A \cup \bar{A} = S$ , mas A e  $\bar{A}$  são mutuamente excludentes, então:

$$P(A \cup \bar{A}) = P(S)$$

$$P(A \cup \bar{A}) = P(A) + P(\bar{A})$$

$$P(A) + P(\bar{A}) = 1$$

logo,

$$P(\bar{A}) = 1 - P(A).$$

Teorema 3 - Se A e B são eventos quaisquer, então:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Prova: Para provar o Teorema 3 devemos transformar  $A \cup B$  em eventos mutuamente excludentes, conforme a FIGURA 6.

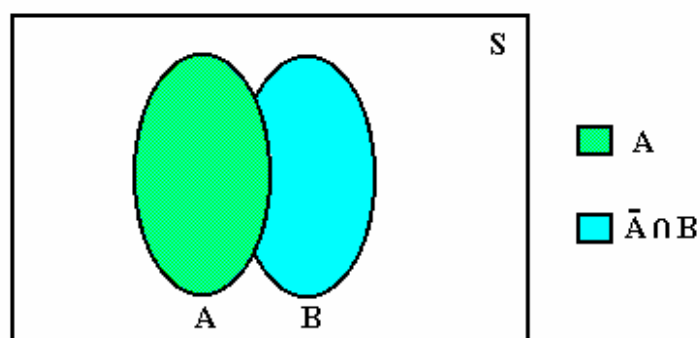


FIGURA 6 - Decomposição de eventos quaisquer em mutuamente excludentes

Tem-se então que:

$$(A \cup B) = A \cup (B \cap \bar{A})$$

e

$$B = (A \cap B) \cup (B \cap \bar{A})$$

logo pela propriedade (c) temos:

$$P(A \cup B) = P[A \cup (B \cap \bar{A})]$$

$$P(A \cup B) = P(A) + P(B \cap \bar{A}) \quad (*)$$

e

$$P(B) = P[(A \cap B) \cup (B \cap \bar{A})]$$

$$P(B) = P(A \cap B) + P(B \cap \bar{A})$$

ou

$$P(B \cap \bar{A}) = P(B) - P(A \cap B) \quad (**)$$

substituindo-se a equação (\*) na equação (\*\*) tem-se:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Decorências do Teorema 3:

Sejam A, B e C eventos quaisquer:

$$P(A \cup B \cup C) = P[(A \cup B) \cup C]$$

$$P(A \cup B \cup C) = P(A \cup B) + P(C) - P[(A \cup B) \cap C]$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P[(A \cap C) \cup (B \cap C)]$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - [P(A \cap C) + P(B \cap C) - P(A \cap B \cap C)]$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Sejam  $A_1, A_2, \dots, A_n$  eventos quaisquer:

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{i=1}^n P(A_i) - \sum_{i < j=2}^n P(A_i \cap A_j) + \sum_{i < j < k=3}^n P(A_i \cap A_j \cap A_k) +$$

$$- \sum_{i < j < k < \ell=4}^n P(A_i \cap A_j \cap A_k \cap A_\ell) + \dots + (-1)^{n-1} P(A_1 \cap A_2 \cap \dots \cap A_n).$$

### RESUMO

$$\begin{array}{l} \cup \\ + \text{ (OU)} \end{array} \left\{ \begin{array}{l} \text{Mutuamente Excludentes} \Rightarrow P(A \cup B) = P(A) + P(B) \\ A \cap B = \emptyset \\ \text{Quaisquer} \Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B) \\ A \cap B \neq \emptyset \end{array} \right.$$

$$\begin{array}{l} \cap \\ \times \text{ (E)} \end{array} \left\{ \begin{array}{l} \text{Independentes} \Rightarrow P(A \cap B) = P(A) \times P(B) \\ A \cap B \neq \emptyset \\ \text{Dependentes} \Rightarrow P(A \cap B) = P(A) \times P(B/A) \end{array} \right.$$

## 5.4 Exercícios

- 1) A probabilidade de 3 jogadores marcarem um penalti é respectivamente:  $\frac{2}{3}$ ;  $\frac{4}{5}$ ;  $\frac{7}{10}$  cobrando uma única vez.
  - a) todos acertarem. ( $\frac{28}{75}$ )
  - b) apenas um acertar. ( $\frac{1}{6}$ )
  - c) todos errarem. ( $\frac{1}{50}$ )
- 2) Numa bolsa com 5 moedas de 1,00 e 10 moedas de 0,50. Qual a probabilidade de ao retirarmos 2 moedas obter a soma 1,50. ( $\frac{10}{21}$ )
- 3) Uma urna contém 5 bolas pretas, 3 vermelhas e 2 brancas. Três bolas são retiradas. Qual a probabilidade de retirar 2 pretas e 1 vermelha ?
  - a) sem reposição ( $\frac{1}{4}$ )
  - b) com reposição ( $\frac{9}{40}$ )
- 4) Numa classe há 10 homens e 20 mulheres, metade dos homens e metade das mulheres possuem olhos castanhos. Ache a probabilidade de uma pessoa escolhida ao acaso ser homem ou ter olhos castanhos. ( $\frac{2}{3}$ )
- 5) A probabilidade de um homem estar vivo daqui a 20 anos é de 0.4 e de sua mulher é de 0.6. Qual a probabilidade de que:
  - a) ambos estejam vivos no período ? (0.24)
  - b) somente o homem estar vivo ? (0.16)
  - c) ao menos a mulher estar viva ? (0.6)
  - d) somente a mulher estar viva? (0.36)
- 6) Faça o exercício anterior considerando 0,5 a chance do homem estar vivo e 0,2 a chance da mulher estar viva e compara os resultados.
- 7) Uma urna contém 5 fichas vermelhas e 4 brancas. Extraem-se sucessivamente duas fichas, sem reposição e constatou-se que a 1ª é branca.
  - a) qual a probabilidade da segunda também ser branca ? ( $\frac{3}{8}$ )
  - b) qual a probabilidade da 2ª ser vermelha ? ( $\frac{5}{8}$ )
- 8) Numa cidade 40 % da população possui cabelos castanhos, 25% olhos castanhos e 15% olhos e cabelos castanhos. Uma pessoa é selecionada aleatoriamente.
  - a) se ela tiver olhos castanhos, qual a probabilidade de também ter cabelos castanhos?
  - b) se ela tiver cabelos castanhos, qual a probabilidade de ter olhos castanhos ? ( $\frac{3}{5}$ ) ( $\frac{3}{8}$ )

9) Um conjunto de 80 pessoas tem as características abaixo:

	BRASILEIRO	ARGENTINO	URUGUAIO	TOTAL
MASCULINO	18	12	10	40
FEMININO	20	05	15	40
TOTAL	38	17	25	80

Se retirarmos uma pessoa ao acaso, qual a probabilidade de que ela seja:

- brasileira ou uruguaia.  $(63/80)$
- do sexo masculino ou tenha nascido na argentina.  $(9/16)$
- brasileiro do sexo masculino.  $(18/80)$
- uruguaio do sexo feminino.  $(15/80)$
- ser mulher se for argentino.  $(5/17)$

10) Um grupo de pessoas está assim formado:

	Médico	Engenheiro	Veterinário
Masc.	21	13	15
Femin.	12	08	17

Escolhendo-se, ao acaso, uma pessoa do grupo, qual a probabilidade de que seja:

- Uma mulher que fez o curso de medicina ?
- Uma pessoa que fez o curso de medicina ?
- Um engenheiro dado que seja homem ?
- Não ser médico dado que não seja homem ?

11) Num ginásio de esportes, 26% dos frequentadores jogam vôlei, 36% jogam basquete e 12% praticam os dois esportes. Um dos frequentadores é sorteado para ganhar uma medalha. Sabendo-se que ele joga basquete, qual a probabilidade de que também jogue vôlei ?

12) A probabilidade de um aluno resolver um determinado problema é de  $1/5$  e a probabilidade de outro é de  $5/6$ . Sabendo que os alunos tentam solucionar o problema independentemente. Qual a probabilidade do problema ser resolvido :

- somente pelo primeiro ?
- ao menos por um dos alunos ?
- por nenhum ?

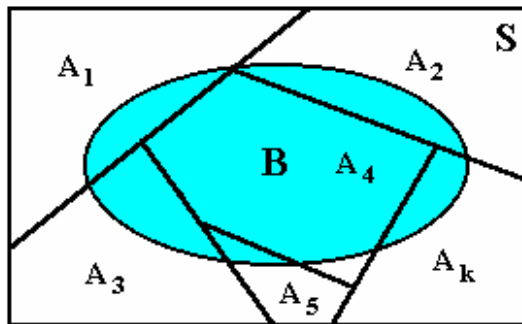
## 5.5 Teorema de Bayes

Definição: Seja  $S$  um espaço amostral e  $A_1, A_2, \dots, A_k$ ,  $k$  eventos. Diz-se que  $A_1, A_2, \dots, A_k$ , formam uma partição de  $S$  se:

$$A_i \neq \emptyset, i = 1, 2, \dots, k$$

$$\bigcup_{i=1}^k A_i = S,$$

$$A_i \cap A_j = \emptyset, i \neq j$$



$A_1, A_2, A_3, A_4, A_5, \dots, A_k$  formam uma partição de  $S$ .

FIGURA 6.7 - Diagrama representativo do Teorema de Bayes

Seja  $B$  um evento qualquer de  $S$ , onde:

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \dots \cup (B \cap A_k)$$

$$P(B) = \sum_{j=1}^k P(A_j) \cdot P\left(\frac{B}{A_j}\right), j = 1, 2, \dots, k \quad (*)$$

$$P(B \cap A_i) = P(A_i) \cdot P\left(\frac{B}{A_i}\right), \quad (**)$$

como



$$P\left(\frac{A_i}{B}\right) = \frac{P(B \cap A_i)}{P(B)}, \quad (***)$$

substituindo as equações (\*) e (\*\*) na equação (\*\*\*) temos:

$$P\left(\frac{A_i}{B}\right) = \frac{P(A_i)P\left(\frac{B}{A_i}\right)}{\sum_{j=1}^k P(A_j)P\left(\frac{B}{A_j}\right)}, j = 1, 2, \dots, k.$$

Exemplo:

Urna		U <sub>1</sub>	U <sub>2</sub>	U <sub>3</sub>
Cores	Azul	3	4	3
	Branca	1	3	3
	Preta	5	2	3

Escolhe-se uma urna ao acaso e dela extrai-se uma bola ao acaso, verificando-se que ela é branca. Qual a probabilidade dela ter saído da urna:

U<sub>1</sub> ?                      U<sub>2</sub> ?                      U<sub>3</sub> ?

2) Temos 2 caixas: na primeira há 3 bolas brancas e 7 pretas e na segunda, 1 branca e 5 pretas. De uma caixa escolhida aleatoriamente, selecionou-se uma bola e verificou-se que é preta. Qual a probabilidade de que tenha saído da primeira caixa ? segunda caixa ?

## 5.6 Variáveis aleatórias

Ao descrever um espaço amostral (S) associado a um experimento (Ω) especifica-se que um resultado individual necessariamente, seja um número. Contudo, em muitas situações experimentais, estaremos interessados na mensuração de alguma coisa e no seu registro como um número.

Definição: Seja (Ω) um experimento aleatório e seja (S) um espaço amostral associado ao experimento. Uma função de X, que associe a cada elemento  $s \in S$  um número real  $x(s)$ , é denominada variável aleatória.

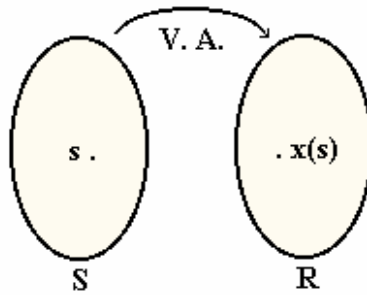


FIGURA 1 - Representação de uma variável aleatória

Uma variável  $X$  será discreta (V.A.D.) se o número de valores de  $x(s)$  for finito ou infinito numerável. Caso encontrarmos  $x(s)$  em forma de intervalo ou um conjunto de intervalos, teremos uma variável aleatória contínua (V.A.C.).

### 5.7 Função de Probabilidade

A probabilidade de que uma variável aleatória " $X$ " assumo o valor " $x$ " é uma função de probabilidade, representada por  $P(X = x)$  ou  $P(x)$ .

#### 5.7.1 Função de Probabilidade de uma V.A.D.

A função de probabilidade para uma variável aleatória discreta é chamada de função de probabilidade no ponto, ou seja, é o conjunto de pares  $(x_i, P(x_i))$ ,  $i = 1, 2, \dots, n, \dots$ , conforme mostra a FIGURA 9.

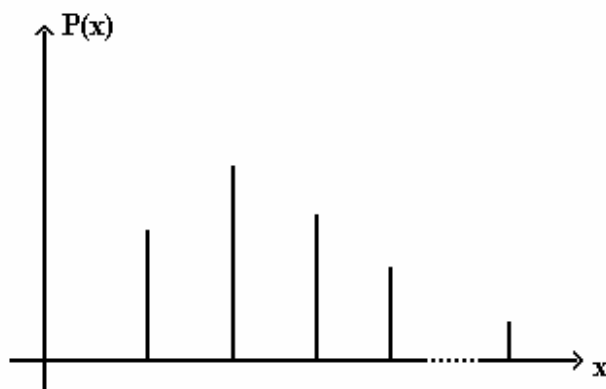


FIGURA 6.1 - Distribuição de probabilidade de uma V.A.D.

Para cada possível resultado de  $x$  teremos:

$$0 \leq P(X) \leq 1;$$

$$\sum_{i=1}^{\infty} P(X_i) = 1$$

### 5.7.2 Função de Repartição para V.A.D.

Seja  $X$  uma variável aleatória discreta.

Define-se Função de Repartição da Variável aleatória  $X$ , no ponto  $x_i$ , como sendo a probabilidade de que  $X$  assumira um valor menor ou igual a  $x_i$ , isto é:

$$F(X) = P(X \leq x_i)$$

#### Propriedades:

$$1a) F(X) = \sum_{x_i \leq x} P(x_i)$$

$$2a) F(-\infty) = 0$$

$$3a) F(+\infty) = 1$$

$$4a) P(a < x \leq b) = F(b) - F(a)$$

$$5a) P(a \leq x \leq b) = F(b) - F(a) + P(X = a)$$

$$6a) P(a < x < b) = F(b) - F(a) - P(X = b)$$

$$7a) F(X) \text{ é contínua à direita} \Rightarrow \lim_{x \rightarrow x_0} F(X) = F(X_0)$$

8a)  $F(X)$  é descontínua à esquerda, nos pontos em que a probabilidade é diferente de zero.

$$\lim_{x \rightarrow x_0} F(X) \neq F(X_0), \text{ para } P(X = x_0) \neq 0$$

9a) A função não é decrescente, isto é,  $F(b) \geq F(a)$  para  $b > a$ .

### 5.6.3 Esperança Matemática de V.A.D.

Definição: Seja  $X$  uma V.A.D., com valores possíveis  $x_1, x_2, \dots, x_n, \dots$ ; Seja  $P(x_i) = P(X = x_i)$ ,  $i = 1, 2, \dots, n, \dots$ . Então, o valor esperado de  $X$  (ou Esperança Matemática de  $X$ ), denotado por  $E(X)$  é definido como

$$E(X) = \sum_{i=1}^{\infty} x_i P(x_i)$$

se a série  $E(X) = \sum_{i=1}^{\infty} x_i P(x_i)$  convergir absolutamente, isto é, se  $\sum_{i=1}^{\infty} |x_i| P(x_i) < \infty$ , este número é também denominado o valor médio de X, ou expectância de X.

#### 5.7.4 Variância de uma V.A.D.

Definição: Seja X uma V.A.D. . Define-se a variância de X, denotada por  $V(X)$  ou  $\sigma_X^2$ , da seguinte maneira:

$$V(X) = \sum_{i=1}^{\infty} (x_i - E(X))^2 \cdot P(x_i) \text{ ou } V(X) = E(X^2) - [E(X)]^2$$

onde  $E(X^2) = \sum_{i=1}^{\infty} x_i^2 P(x_i)$  e a raiz quadrada positiva de  $V(X)$  é denominada o desvio-padrão de X, e denotado por  $\sigma_X$ .

#### 5.7.5 Função de Probabilidade de uma V.A.C.

No instante em que X é definida sobre um espaço amostral contínuo, a função de probabilidade será contínua, onde a curva limitada pela área em relação ao valores de x será igual a 1.

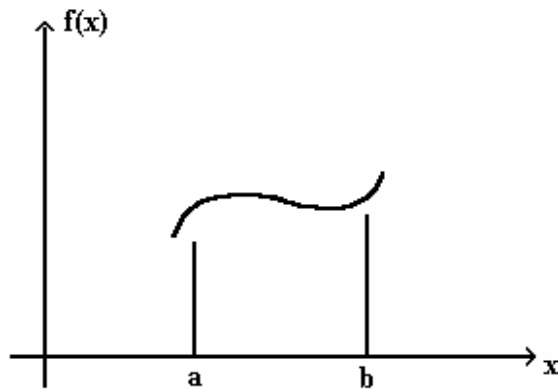


FIGURA 6.2 - Distribuição de probabilidade de uma V.A.C.

Se quisermos calcular a probabilidade de X assumir um valor x entre "a" e "b" devemos calcular:

$$P(a \leq x \leq b) = \int_a^b f(x) dx$$

Pelo fato de que a área representa probabilidade, e a mesma tem valores numéricos positivos, logo a função precisa estar inteiramente acima do eixo das abscissas (x).

Definição: A função  $f(x)$  é uma Função Densidade de Probabilidade (f.d.p.) para uma V.A.C.  $X$ , definida nos reais quando

$$f(x) \geq 0;$$

$$\int_{-\infty}^{+\infty} f(x) dx = 1;$$

$$P(a \leq x \leq b) = \int_a^b f(x) dx .$$

### 5.7.6 Função de Repartição para V.A.C.

Seja  $X$  uma variável aleatória contínua.

Define-se Função de Repartição da Variável aleatória  $X$ , no ponto  $x_i$ , como sendo:

$$F(X) = \int_{-\infty}^x f(x) dx$$

$$\equiv P(a \leq x \leq b) = P(a < x < b) = P(a < x \leq b) = P(a \leq x < b)$$

### 5.7.7 Esperança Matemática de uma V.A.C.

Definição: Seja  $X$  uma V.A.C. com f.d.p.  $f(x)$ . O valor esperado de  $X$  é definido como

$$E(X) = \int_{-\infty}^{+\infty} x.f(x) dx$$

pode acontecer que esta integral imprópria não convirja.

Conseqüentemente, diremos que  $E(X)$  existirá se, e somente se,  $\int_{-\infty}^{+\infty} |x| f(x)$  for finita.

### 5.7.8 Variância de uma V.A.C.

Definição: Seja  $X$  uma V.A.C. de uma função distribuição de probabilidade (f.d.p.). A variância de  $X$  é:

$$V(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x) dx \text{ ou } V(X) = E(X^2) - [E(X)]^2$$

onde

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx$$

### 5.8 Exemplos

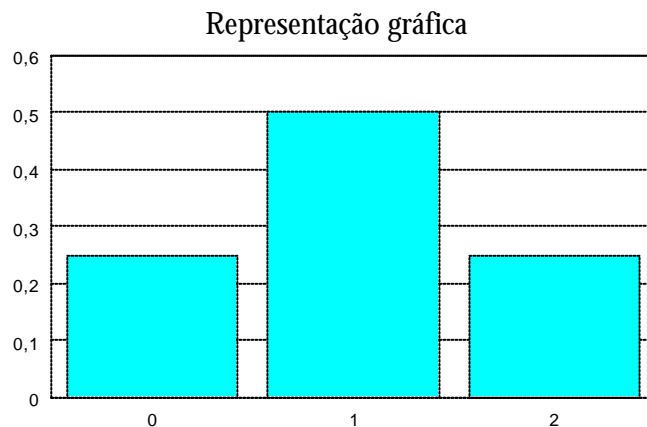
#### - Variável Aleatória Discreta

Seja  $X$  o lançamento de duas moedas e descrever o experimento em função da obtenção do número de caras:

- i) determinar a função de probabilidade e represente graficamente;
- ii) construir a função de repartição e represente graficamente;
- iii) Use as propriedades para determinar:
  - a)  $P(0 < x < 2)$ ; b)  $P(0 \leq x \leq 1)$ ; c)  $P(0 < x \leq 2)$ ; d)  $F(1)$ ; e)  $F(2)$
- iv)  $E(X)$  e  $V(X)$

i)

$X_i$	$P(X = x_i)$
0	1/4
1	2/4
2	1/4
$\Sigma$	4/4



ii)

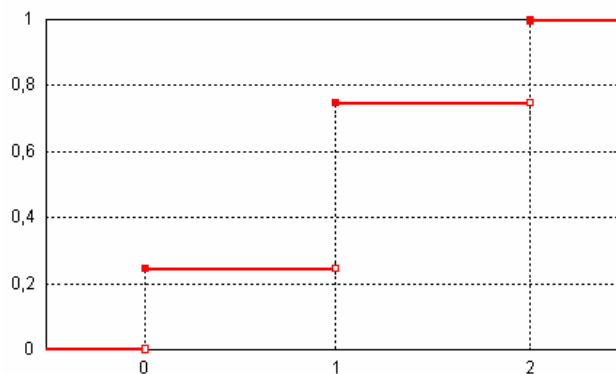
$$F(x) = 0 \text{ se } x < 0$$

$$F(X) = 1/4 \text{ se } 0 \leq x < 1$$

$$F(X) = 3/4 \text{ se } 1 \leq x < 2$$

$$F(X) = 1 \text{ se } x \geq 2$$

Representação gráfica



iii)

$$a) P(0 < x < 2) = F(2) - F(0) - P(X = 2) = 1 - \frac{1}{4} - \frac{1}{4} = \frac{3}{4}$$

$$b) P(0 \leq x \leq 1) = F(1) - F(0) + P(X = 0) = \frac{3}{4} - \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$$

$$c) P(0 < x \leq 2) = F(2) - F(0) = 1 - \frac{1}{4} = \frac{3}{4}$$

$$d) F(1) = 3/4 \quad e) F(2) = 1$$

iv) Esperança Matemática

$$E(X) = \sum_{i=1}^{\infty} x_i P(x_i) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{2}{4} + 2 \cdot \frac{1}{4} = 1$$

Variância

$$E(X^2) = \sum_{i=1}^{\infty} x_i^2 P(x_i) = 0^2 \cdot \frac{1}{4} + 1^2 \cdot \frac{2}{4} + 2^2 \cdot \frac{1}{4} = \frac{6}{4}$$

$$V(X) = E(X^2) - [E(X)]^2 = \frac{6}{4} - \frac{4}{4} = \frac{2}{4} = 0,5$$

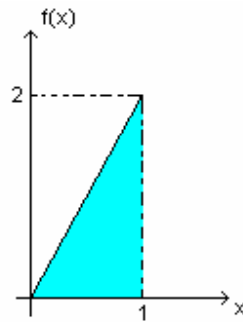
## - Variável Aleatória Contínua

Seja  $X$  uma variável aleatória contínua:

$$f(x) = \begin{cases} 2x, & \text{para } 0 < x < 1 \\ 0, & \text{para qualquer outro valor} \end{cases}$$

- i) represente graficamente função densidade de probabilidade;
- ii) determinar a função de repartição e represente graficamente;
- iii) Determine  $P\left(\frac{1}{4} \leq x \leq \frac{3}{4}\right)$  e  $P\left(\frac{1}{4} \leq x < 1\right)$
- iv)  $E(X)$  e  $V(X)$

i)



ii)

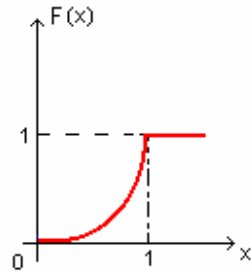
$$\text{para } x < 0 \quad \Rightarrow F(X) = \int_{-\infty}^x 0 \, dx = 0$$

$$\text{para } 0 \leq x < 1 \quad \Rightarrow F(X) = \int_{-\infty}^x 0 \, dx + \int_0^x 2x \, dx = 0 + [x^2]_0^x = x^2$$

$$\text{para } x \geq 1 \quad \Rightarrow F(X) = \int_{-\infty}^x 0 \, dx + \int_0^1 2x \, dx + \int_1^{+\infty} 0 \, dx = 0 + [x^2]_0^1 + 0 = 1$$



### Representação gráfica



$$\text{iii) } P\left(\frac{1}{4} \leq x \leq \frac{3}{4}\right) = \int_{\frac{1}{4}}^{\frac{3}{4}} 2x \, dx = \left[x^2\right]_{\frac{1}{4}}^{\frac{3}{4}} = \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = \frac{8}{16} = 0,5$$

$$\text{iv) } E(X) = \int_0^1 x f(x) \, dx = \int_0^1 x \cdot 2x \, dx = \int_0^1 2x^2 f(x) \, dx = \frac{2}{3} \left[x^3\right]_0^1 = \frac{2}{3}$$

$$E(X^2) = \int_0^1 x^2 f(x) \, dx = \int_0^1 x^2 \cdot 2x \, dx = \int_0^1 2x^3 f(x) \, dx = \frac{2}{4} \left[x^4\right]_0^1 = \frac{1}{2}$$

logo,

$$V(X) = E(X^2) - [E(X)]^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{9-8}{18} = \frac{1}{18}$$

### 5.9 Exercícios

1) Admita que a variável X tome valores 1, 2 e 3 com probabilidades  $1/3$ ,  $1/6$  e  $1/2$  respectivamente.

a) Determine sua função de repartição e represente graficamente.

b) Calcule usando as propriedades:

b.1) a)  $P(1 < x < 3)$ ; b)  $P(1 \leq x \leq 2)$ ; c)  $P(1 < x \leq 3)$ ; d)  $F(1)$ ; e)  $F(2)$

c)  $E(X)$  e  $V(X)$

2) No lançamento simultâneo de dois dados, considere as seguintes variáveis aleatórias:

$X$  = número de pontos obtidos no 1º dado.

$Y$  = número de pontos obtidos no 2º dado.

a) Construir a distribuição de probabilidade através de uma tabela e gráfico das seguinte variáveis:

i)  $W = X - Y$

ii)  $A = 2 Y$

iii)  $Z = X \cdot Y$

b) Construir a função de repartição das Variáveis  $W$ ,  $A$  e  $Z$

c) Aplicar as propriedades e determinar:

i)  $P(-3 < W \leq 3)$

v)  $P(Z = 3)$

ii)  $P(0 \leq W \leq 4)$

vi)  $P(A \geq 11)$

iii)  $P(A > 6)$

vii)  $P(20 \leq Z \leq 35)$

iv)  $P(Z \leq 5.5)$

viii)  $P(3,5 < Z < 34)$

d) Determine  $E(W)$ ,  $E(A)$ ,  $E(Z)$ ,  $V(W)$ ,  $V(A)$  e  $V(Z)$

3) Uma variável aleatória discreta tem a distribuição de probabilidade dada por:

$$P(X) = \frac{K}{x} \text{ para } x = 1, 3, 5 \text{ e } 7$$

a) calcule o valor de  $k$

b) Calcular  $P(X=5)$

c)  $E(X)$

d)  $V(X)$

4) Seja  $Z$  a variável aleatória correspondente ao número de pontos de uma peça de dominó.

a) Construir a tabela e traçar o gráfico  $P(Z)$ .

b) Determinar  $F(Z)$  e traçar o gráfico.

c) Calcular  $P(2 \leq Z < 6)$ .

d) Calcular  $F(8)$ .

e)  $E(Z)$  e  $V(Z)$ .

5) Seja  $f(x) = \begin{cases} \frac{3}{2}(1-x^2), & 0 < x < 1, \\ 0, & \text{caso contrário} \end{cases}$ ,

- i) Ache a função de repartição e esboce o gráfico.
- ii) Determine  $E(X)$  e  $V(X)$ .

6) Seja  $f(x) = \begin{cases} \frac{1}{2}x, & 0 \leq x \leq 2, \\ 0, & \text{caso contrário} \end{cases}$ ,

- i) Ache a função de repartição e esboce o gráfico.
- ii)  $P(1 < x < 1,5)$ .
- iii)  $E(X)$  e  $V(X)$ .

7) Uma variável aleatória  $X$  tem a seguinte f.d.p.:

$x < 0$	$f(x) = 0$
$0 \leq x < 2$	$f(x) = k$
$2 \leq x < 4$	$f(x) = k(x - 1)$
$x \geq 4$	$f(x) = 0$

- a) Represente graficamente  $f(x)$ .
- b) Determine  $k$ .
- c) Determine  $F(X)$  e faça o gráfico
- d)  $E(X)$  e  $V(X)$

8) A função de probabilidade de uma V.A.C.  $X$  é  $f(x) = \begin{cases} 6x(1-x), & 0 < x < 1 \\ 0, & \text{caso contrário} \end{cases}$

- a) Determine  $F(X)$  e represente graficamente.
- b) Calcule  $P\left(x \leq \frac{1}{2}\right)$
- c)  $E(X)$  e  $V(X)$

9) Uma variável aleatória  $X$  tem a seguinte f.d.p.:

$$f(x) = 0 \quad x < 0$$

$$f(x) = Ax \quad 0 \leq x < 500$$

$$f(x) = A(100 - x) \quad 500 \leq x < 1000$$

$$f(x) = 0 \quad x \geq 1000$$

a) Determine o valor de  $A$ .

b)  $P(250 \leq x \leq 750)$ .

10) Dada a função de repartição:

$$F(X) = 0 \quad \text{para } x < -1$$

$$F(X) = \frac{x+1}{2} \quad \text{para } -1 \leq x < 1$$

$$F(X) = 1 \quad \text{para } x \geq 1$$

a) Calcule:  $P\left(-\frac{1}{2} \leq x \leq \frac{1}{2}\right)$ , b)  $P(X = 0)$

# 6 Distribuições de Probabilidade

Após termos visto as definições de V.A.D. e V.A.C., citaremos as principais distribuições de probabilidade relacionadas a estas variáveis.

## 6.1 Distribuições Discretas de Probabilidade

### 6.1.1 Distribuição de Bernoulli

Consideramos uma única tentativa de um experimento aleatório podemos ter sucesso ou fracasso nessa tentativa. Seja "p" a probabilidade de sucesso e "q" a probabilidade de fracasso, onde  $p + q = 1$ .

Seja X o número de sucesso em uma única tentativa, logo X pode assumir:

$X = 0$  se ocorrer fracasso e  $X = 1$  se ocorrer sucesso,

ou ainda:

$$X = \begin{cases} 0, & \text{se for fracasso, com } P(X = 0) = q \\ 1, & \text{se for sucesso, com } P(X = 1) = p. \end{cases}$$

Supondo que em uma única tentativas o número de casos possíveis sejam:

$$\underbrace{A, A, \dots, A}_x \text{ sucessos}, \quad \underbrace{\bar{A}, \bar{A}, \dots, \bar{A}}_{n-x} \text{ fracassos}$$

sendo os resultados independentes, temos:

$$P(A \cap A \cap \dots \cap A \cap \bar{A} \cap \bar{A} \cap \dots \cap \bar{A}) = P(A).P(A). \dots . P(A).P(\bar{A}).P(\bar{A}). \dots . P(\bar{A})$$
$$\underbrace{P(A).P(A). \dots . P(A)}_x \text{ sucessos} \cdot \underbrace{P(\bar{A}).P(\bar{A}). \dots . P(\bar{A})}_{n-x} \text{ fracassos}, \text{ onde } P(A) = p \text{ e } P(\bar{A}) = q,$$

nessa condição a variável aleatória  $X$  tem distribuição de Bernoulli, e sua função de probabilidade (f. p.) (função de probabilidade) é dada por:

$$P(X = x) = p^x \cdot q^{n-x}.$$

### 6.1.1.1 Esperança Matemática da Distribuição de Bernoulli

$$E(X) = \sum_{i=1}^{\infty} x_i P(x_i)$$

$$E(X) = x_1 P(x_1) + x_2 P(x_2)$$

$$E(X) = 0 \cdot P(0) + 1 \cdot P(1)$$

$$E(X) = 0 \cdot q + 1 \cdot p$$

$$E(X) = p$$

### 6.1.1.2 Variância da Distribuição de Bernoulli

$$V(X) = E(X^2) - [E(X)]^2$$

onde

$$E(X^2) = \sum_{i=1}^{\infty} x_i^2 P(x_i)$$

$$E(X^2) = x_1^2 P(x_1) + x_2^2 P(x_2)$$

$$E(X^2) = 0 \cdot q + 1 \cdot p = p$$

logo

$$V(X) = p - p^2$$

$$V(X) = p \cdot (1 - p)$$

$$V(X) = p \cdot q$$

### 6.1.2 Distribuição Binomial

O termo "Binomial" é utilizado quando uma variável aleatória esta agrupada em duas classes ou categorias. As categorias devem ser mutuamente excludentes, de modo a deixar bem claro a qual categoria pertence determinada observação; e as classes devem ser coletivamente exaustivas, de forma que nenhum outro resultado fora delas é possível

Sejam, "p" probabilidades de sucesso e "q" probabilidades de falha, ou seja  $p + q = 1$ .

A probabilidade de x sucessos em x tentativas é dado por  $p^x$  e de (n - x) falhas em (n - x) tentativas é dado por  $q^{n-x}$ , onde o número de vezes em que pode ocorrer x sucessos e (n-x) falhas é dado por:

$$C_{n,x} = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

logo, a probabilidade de ocorrer x sucessos com n tentativas será

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

Propriedades necessárias para haver uma utilização da Distribuição Binomial:

- 1a) Número de tentativas fixas;
- 2a) Cada tentativa deve resultar numa falha ou sucesso;
- 3a) As probabilidades de sucesso devem ser iguais para todas as tentativas;
- 4a) Todas as tentativas devem ser independentes.

### 6.1.2.1 Esperança Matemática de Distribuição Binomial

$$E(X) = \sum_{i=1}^{\infty} x_i P(x_i)$$

$$E(X) = x \cdot \binom{n}{x} p^x q^{n-x}$$

como P(X) segue uma Distribuição de Probabilidade, temos:

$$\sum_{i=1}^{\infty} P(X = x_i) = 1 \text{ ou } \sum_{x=1}^{\infty} \binom{n}{x} p^x q^{n-x} = 1$$

logo,

$$E(X) = x \cdot \frac{n!}{x(x-1)!(n-x)!} \sum_{x=1}^{\infty} p^x q^{n-x}$$

$$E(X) = n \cdot \frac{(n-1)!}{(x-1)!(n-x)!} \sum_{x=1}^{\infty} p^x q^{n-x}, \text{ ou seja para } s = x - 1 \text{ e } x = s + 1, \text{ temos}$$

$$E(X) = n \cdot \sum_{x=1}^{\infty} \binom{n-1}{s} p^{s+1} q^{n-(s+1)}$$

$$E(X) = n \cdot \sum_{x=1}^{\infty} \binom{n-1}{s} p^s \cdot p q^{(n-1)-s}$$

$$E(X) = n \cdot p \underbrace{\sum_{x=1}^{\infty} \binom{n-1}{s} p^s \cdot q^{(n-1)-s}}_1$$

$$E(X) = n \cdot p$$

### 6.1.2.2 Variância de uma Distribuição Binomial

$$V(X) = E(X^2) - [E(X)]^2$$

onde

$$E(X^2) = \sum_{i=1}^{\infty} x_i^2 P(x_i)$$

$$E(X^2) = x \cdot x \frac{n!}{x \cdot (x-1)! (n-x)!} \sum_{x=1}^{\infty} p^x q^{n-x}, \text{ para } s = x-1 \text{ e } x = s+1, \text{ temos:}$$

$$E(X^2) = (s+1) \frac{n!}{s! (n-(s+1))!} \sum_{x=1}^{\infty} p^{s+1} q^{n-(s+1)}$$

$$E(X^2) = s \cdot n \sum_{x=1}^{\infty} \frac{(n-1)!}{s! (n-1-s)!} p^s \cdot p q^{n-1-s} + n \sum_{x=1}^{\infty} \frac{(n-1)!}{s! (n-1-s)!} p^s \cdot p q^{n-1-s}$$

$$E(X^2) = n \cdot p \left[ \underbrace{s \binom{n-1}{s} \sum_{x=1}^{\infty} p^s \cdot p q^{n-1-s}}_{(n-1)p} + \underbrace{\sum_{x=1}^{\infty} \binom{n-1}{s} p^s q^{n-1-s}}_1 \right]$$

$$E(X^2) = n \cdot p [(n-1)p + 1]$$

$$E(X^2) = n \cdot p [np - p + 1]$$

$$E(X^2) = n^2 p^2 - np^2 + np$$

$$V(X) = n^2 p^2 - np^2 + np - (np)^2$$

$$V(X) = n \cdot p \cdot (1-p)$$

$$V(X) = n \cdot p \cdot q$$



### 6.1.2.3 Esperança Matemática e Variância para a Probabilidade de Sucesso (p) de uma Distribuição Binomial

$$E(p) = p \text{ e } V(p) = \frac{p \cdot q}{n}$$

### 6.1.3 Distribuição de Poisson

Quando numa distribuição binomial o tamanho "n" das observações for muito grande e a probabilidade "p" de sucesso for muito pequena, a probabilidade x de ocorrência de um determinado número de observações segue uma Distribuição de Poisson.

A aplicação da distribuição segue algumas restrições:

- Somente a chance afeta o aparecimento do evento, contando-se apenas com a sua ocorrência, ou seja, a probabilidade de sucesso "p".

- Uma vez não conhecido o número total de eventos, a distribuição não pode ser aplicada.

Se  $n \rightarrow \infty$  em consequência  $n \gg x$  assim,

$$\frac{n!}{(n-x)!} = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-x+1) \approx n^x$$

logo,

$$P(x) = \frac{n^x}{x!} p^x q^{n-x}.$$

Se  $p \rightarrow 0$  e  $n \gg x$  logo,  $q^{n-x} \approx q^n = (1-p)^n$

assim,

$$P(x) = \frac{(n \cdot p)^x (1-p)^n}{x!}$$

$$P(x) = \frac{(n \cdot p)^x}{x!} \left[ 1 - n \cdot p + \frac{n(n-1)}{2 \cdot 1} (-p)^2 + \dots \right]$$

$$P(x) \cong \frac{(n \cdot p)^x}{x!} \left[ 1 - n \cdot p + \frac{(n \cdot p)^2}{2!} + \dots \right]$$

$$P(x) = \frac{(n.p)^x e^{-n.p}}{x!}$$

Substituindo o valor esperado  $n.p$  por  $\lambda$  e considerando-o como sendo o número médio de ocorrência expresso em unidades de tempo, pode-se dizer que  $\lambda$  é a taxa média de falhas (falha / unid. tempo) e  $t$  o tempo, logo o número médio de falhas será  $\lambda t$ , assim,

$$P(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

fornece a probabilidade de  $x$  falhas no período de tempo  $t$ .

A probabilidade de zero falhas no tempo  $t$  é a confiabilidade do componente em função do tempo.

$$P(0) = R(t) = e^{-\lambda t}$$

### 6.1.3.1 Esperança Matemática da Distribuição de Poisson

$$E(X) = \sum_{i=1}^{\infty} x_i P(x_i)$$

$$E(X) = \sum_{x=1}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!}$$

$$E(X) = \sum_{x=1}^{\infty} \frac{e^{-\lambda} \lambda^x}{(x-1)!}, \text{ substituindo } s = x - 1 \text{ e } x = s + 1 \text{ temos:}$$

$$E(X) = \sum_{x=1}^{\infty} \frac{e^{-\lambda} \lambda^{s+1}}{s!}$$

$$E(X) = \lambda \underbrace{\sum_{x=1}^{\infty} \frac{e^{-\lambda} \lambda^s}{s!}}_1$$

$$E(x) = \lambda$$

### 6.1.3.2 Variância da Distribuição de Poisson

$$V(X) = E(X^2) - [E(X)]^2$$

onde

$$E(X^2) = \sum_{i=1}^{\infty} x_i^2 P(x_i)$$

$$E(X^2) = \sum_{x=1}^{\infty} x \cdot x \frac{e^{-\lambda} \lambda^x}{x(x-1)!}, \text{ substituindo } s = x - 1 \text{ e } x = s + 1 \text{ temos:}$$

$$E(X^2) = \sum_{s=1}^{\infty} (s+1) \frac{e^{-\lambda} \lambda^{s+1}}{s!}$$

$$E(X^2) = \lambda \sum_{s=1}^{\infty} (s+1) \frac{e^{-\lambda} \lambda^s}{s!}$$

$$E(X^2) = \lambda \left[ \underbrace{\sum_{s=1}^{\infty} s \frac{e^{-\lambda} \lambda^s}{s!}}_{E(X)} + \underbrace{\sum_{s=1}^{\infty} \frac{e^{-\lambda} \lambda^s}{s!}}_1 \right]$$

$$E(X)^2 = \lambda^2 + \lambda$$

$$V(X) = \lambda^2 + \lambda - (\lambda)^2$$

$$V(X) = \lambda$$

## 6.2 Exercícios

- 1) Admitindo-se o nascimento de meninos e meninas sejam iguais, calcular a probabilidade de um casal com 6 filhos ter:
  - a) 4 filhos e 2 filhas
  - b) 3 filhos e 3 filhas
  
- 2) Em 320 famílias com 4 crianças cada uma, quantas se esperaria que tivessem:
  - a) nenhuma menina;
  - b) 3 meninos
  - c) 4 meninos
  
- 3) Um time X tem  $\frac{2}{3}$  de probabilidade de vitória sempre que joga. Se X jogar 5 partidas, calcule a probabilidade de:
  - a) X vencer exatamente 3 partidas;
  - b) X vencer ao menos uma partida;
  - c) X vencer mais da metade das partidas;
  - d) X perder todas as partidas;
  
- 4) A probabilidade de um atirador acertar um alvo é  $\frac{1}{3}$ . Se ele atirar 6 vezes, qual a probabilidade de:
  - a) acertar exatamente 2 tiros;
  - b) não acertar nenhum tiro.
  
- 5) Num teste de certo-errado, com 100 perguntas, qual a probabilidade de um aluno, respondendo as questões ao acaso, acertar 70% das perguntas ?

- 6) Se 5% das lâmpadas de certa marca são defeituosas, achar a probabilidade de que, numa amostra de 100 lâmpadas, escolhidas ao acaso, tenhamos:
- nenhuma defeituosa (use binomial e poisson)
  - 3 defeituosas;
  - mais do que uma boa;
- 7) Uma fabrica de pneus verificou que ao testar seus pneus nas pistas, havia em média um estouro de pneu a cada 5.000 km.
- qual a probabilidade que num teste de 3.000 km haja no máximo um pneu estourado ?
  - Qual a probabilidade de um carro andar 8.000 km sem estourar nenhum pneu ?
- 8) Certo posto de bombeiros recebe em média 3 chamadas por dia. Calcular a probabilidade de:
- receber 4 chamadas num dia;
  - receber 3 ou mais chamadas num dia;
  - 22 chamadas numa semana.
- 9) A média de chamadas telefônicas em uma hora é 3. Qual a probabilidade:
- receber exatamente 3 chamadas numa hora;
  - receber 4 ou mais chamadas em 90 minutos;
  - 75 chamadas num dia;
- 10) Na pintura de paredes aparecem defeitos em média na proporção de 1 defeito por metro quadrado. Qual a probabilidade de aparecerem 3 defeitos numa parede 2 x 2 m?
- 11) Suponha que haja em média 2 suicídios por ano numa população de 50.000 hab. Em uma cidade de 100.000 habitantes, encontre a probabilidade de que um dado ano tenha havido: a) nenhum suicídio; b) 1 suicídio; c) 2 ou mais suicídios.
- 12) Suponha 400 erros de impressão distribuídos aleatoriamente em um livro de 500 páginas. Encontre a probabilidade de que uma dada página contenha:
- nenhum erro;
  - 100 erros em 200 páginas.

## 6.3 Distribuições Contínuas de Probabilidade

### 6.3.1 Distribuição Uniforme

É uma distribuição de probabilidade usada para variáveis aleatórias contínuas, definida num intervalo  $[a, b]$ , e sua função densidade de probabilidade é dada por:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{se } a \leq x \leq b \\ 0 & \text{se } x < a \text{ ou } x > b \end{cases}.$$

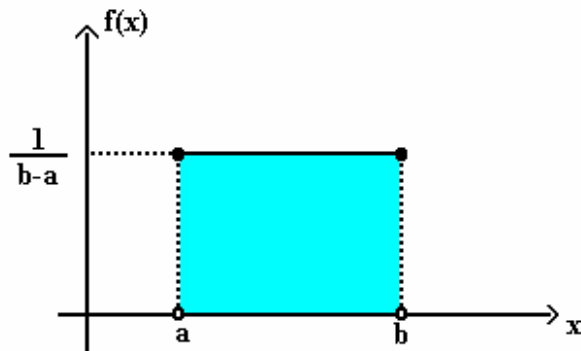


FIGURA 6.1 - Representação de uma Distribuição Uniforme

#### 6.3.1.1 Esperança Matemática da Distribuição Uniforme

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

$$E(X) = \int_a^b x \frac{1}{b-a} dx$$

$$E(X) = \frac{1}{b-a} \left[ \frac{x^2}{2} \right]_a^b$$

$$E(X) = \frac{b^2 - a^2}{2(b-a)}$$

$$E(X) = \frac{(b-a)(b+a)}{2(b-a)}$$

$$E(X) = \frac{(b+a)}{2}$$

### 6.3.1.2 Variância da Distribuição Uniforme

$$V(X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx$$

$$E(X^2) = \int_a^b x^2 \frac{1}{b-a} dx$$

$$E(X^2) = \frac{1}{b-a} \int_a^b x^2 dx$$

$$E(X^2) = \frac{1}{b-a} \left[ \frac{x^3}{3} \right]_a^b$$

$$E(X^2) = \frac{b^3 - a^3}{3(b-a)}$$

$$E(X^2) = \frac{(b-a)(b^2 + ab + a^2)}{3(b-a)}$$

$$E(X^2) = \frac{b^2 + ab + a^2}{3}$$

$$V(X) = \frac{b^2 + ab + a^2}{3} - \left( \frac{b+a}{2} \right)^2$$

$$V(X) = \frac{b^2 + ab + a^2}{3} - \frac{b^2 - ab + a^2}{4}$$

$$V(X) = \frac{4b^2 + 4ab + 4a^2 - 3b^2 - 6ab - 3a^2}{12}$$

$$V(X) = \frac{b^2 - 2ab + a^2}{12}$$

$$V(X) = \frac{(b-a)^2}{12}$$

### 6.3.2 Distribuição Normal ou Gaussiana

É um modelo de distribuição contínua de probabilidade, usado tanto para variáveis aleatórias discretas como contínuas.

Uma variável aleatória  $X$ , que tome todos os valores reais  $-\infty < x < +\infty$  tem distribuição normal quando sua função densidade de probabilidade (f.d.p.) for da forma:

$$f(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}, -\infty < x < +\infty$$

Os parâmetros  $\mu$  e  $\sigma$  seguem as seguintes condições:

$$-\infty < \mu < +\infty \text{ e } \sigma > 0.$$

### 6.3.2.1 Propriedades da Distribuição Normal

a)  $f$  é uma f.d.p. legítima para  $f(x) \geq 0$ , logo

$$\int_{-\infty}^{+\infty} f(x) d(x) = 1$$

fazendo  $t = \frac{x - \mu}{\sigma}$  temos:

$$\int_{-\infty}^{+\infty} f(x) d(x) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{t^2}{2}} dt = I.$$

Para calcular esta integral usaremos um artifício, ou seja, no lugar de  $I$  usaremos  $I^2$ .

$$I^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt \cdot \int_{-\infty}^{+\infty} e^{-\frac{s^2}{2}} ds$$

$$I^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{(t^2+s^2)}{2}} dt ds$$

introduzindo coordenadas polares para realizar o cálculo dessa integral dupla, temos:

$$s = r \cos\alpha \text{ e } t = r \sin\alpha$$

conseqüentemente o elemento de área  $ds dt$  se torna  $r dr d\alpha$ .

Como  $-\infty < s < +\infty$  e  $-\infty < t < +\infty$ ,  $0 < r < +\infty$  e  $0 < \alpha < 2\pi$ , portanto

$$I^2 = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{+\infty} r e^{-\frac{r^2}{2}} dr d\alpha$$

$$I^2 = \frac{1}{2\pi} \int_0^{2\pi} e^{-\frac{r^2}{2}} \Big|_0^{+\infty} d\alpha$$

$$I^2 = \frac{1}{2\pi} \int_0^{2\pi} -(0-1) d\alpha$$

$$I^2 = \frac{1}{2\pi} \alpha \Big|_0^{2\pi}$$

$$I^2 = \frac{1}{2\pi} 2\pi = 1$$

$I^2 = 1$ , logo  $I = 1$  como queríamos mostrar.

b) O aspecto gráfico da função  $f$  tem:

- Semelhança de um sino, unimodal e simétrico em relação a média  $\mu$ .
- A especificação da média  $\mu$  e do desvio padrão  $\sigma$  é completamente evidenciado.
- A área total da curva equivale a 100%.
- A área total da curva equivale a 100%.

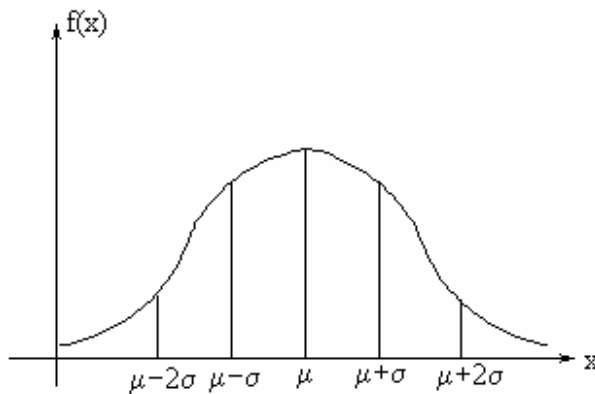


FIGURA 7.2 - Distribuição Normal em função da  $\mu$  e  $\sigma$

### 6.3.2.2 Esperança Matemática da Distribuição Normal

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

$$E(X) = \int_{-\infty}^{+\infty} x \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2} dx$$

fazendo  $z = \frac{x-\mu}{\sigma}$ ,  $\partial z = \frac{\partial x}{\sigma}$  e  $x = z \sigma + \mu$ ,

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma z + \mu) e^{-\frac{z^2}{2}} dz$$

$$E(X) = \frac{1}{\sqrt{2\pi}} \sigma \underbrace{\int_{-\infty}^{+\infty} z e^{-\frac{z^2}{2}} dz}_{\text{zero}} + \mu \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz}_{\text{um}}$$



$$E(X) = \mu$$

### 6.3.2.3 Variância da Distribuição Normal

$$V(X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx$$

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

fazendo  $z = \frac{x-\mu}{\sigma}$ ,  $dz = \frac{dx}{\sigma}$  e  $x = z\sigma + \mu$ ,

$$E(X^2) = \frac{1}{\sigma\sqrt{2\pi}} \sigma \int_{-\infty}^{+\infty} (\sigma z + \mu)^2 e^{-\frac{z^2}{2}} dz$$

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \sigma^2 \int_{-\infty}^{+\infty} z^2 e^{-\frac{z^2}{2}} dz + \frac{1}{\sqrt{2\pi}} 2\mu \sigma \underbrace{\int_{-\infty}^{+\infty} z e^{-\frac{z^2}{2}} dz}_{\text{zero}} + \mu^2 \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz}_{\text{um}}$$

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \sigma^2 \int_{-\infty}^{+\infty} z^2 e^{-\frac{z^2}{2}} dz + \mu^2$$

integrando por partes temos que  $\int u dv = u dv - \int v du$

$$z^2 = u \quad dv = e^{-\frac{z^2}{2}} dz$$

$$du = 2z dz \quad v = -e^{-\frac{z^2}{2}}$$

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \sigma^2 \left[ z^2 e^{-\frac{z^2}{2}} \right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} -e^{-\frac{z^2}{2}} dz + \mu^2$$

$$E(X^2) = 0 + \sigma^2 \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz}_{\text{um}} + \mu^2$$

$$E(X^2) = \sigma^2 + \mu^2$$

logo,

$$V(X) = \sigma^2$$

### 6.3.2.4 Distribuição Normal Padronizada

Tem como objetivo solucionar a complexidade da  $f(x)$  através da mudança de variável.  $f(z)$ .

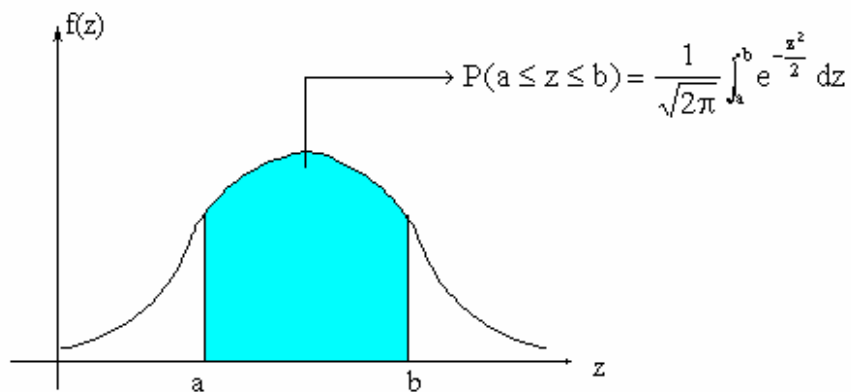


FIGURA 7.4 - Complemento da Distribuição Normal Padronizada

Fazendo  $z = \frac{x - \mu}{\sigma}$  e  $z \sim N(0,1)$  temos que

$$f(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}},$$

com  $E(z) = 0$  e  $VAR(z) = 1$ .

onde:

$z$  = número de desvios padrões a contar da média

$x$  = valor arbitrário

$\mu$  = média da distribuição normal

$\sigma$  = desvio padrão da distribuição normal

Estas probabilidades estão tabeladas e este caso particular é chamado de Forma Padrão da Distribuição Normal.

### 6.3.3 Distribuição "t" de Student

Trata-se de um modelo de distribuição contínua que se assemelha à distribuição normal padrão,  $N \sim (0,1)$ . É utilizada para inferências estatísticas, particularmente, quando se tem amostras com tamanhos inferiores a 30 elementos.

A distribuição t também possui parâmetros denominado "grau de liberdade -  $\varphi$ ". A média da distribuição é zero e sua variância é dada por:

$$\text{VAR}[t_{\varphi}] = \sigma^2(t_{\varphi}) = \frac{\varphi}{\varphi - 2}, \text{ para } \varphi > 2.$$

A distribuição t é simétrica em relação a sua média.

### 6.4 Exercícios

- 1) As alturas dos alunos de uma determinada escola são normalmente distribuídas com média 1,60 m e desvio padrão 0,30 m. Encontre a probabilidade de um aluno escolhido ao acaso medir:
  - a) entre 1,50 e 1,80 m
  - b) mais que 1,75 m
  - c) menos que 1,48 m
  - d) entre 1,54 e 1,58 m
  - e) menos que 1,70 m
  - f) exatamente 1,83 m
  
- 2) A duração de certo componente eletrônico tem média 850 dias e desvio padrão 45 dias. Qual a probabilidade do componente durar:
  - a) entre 700 e 1000 dias
  - b) menos que 750 dias
  - c) mais que 850 dias
- d) Qual deve ser o número de dias necessários para que tenhamos de repor 5% dos componentes. (R = 776 dias)
  
- 3) Um produto pesa, em média, 10 g, com desvio padrão de 2 g. É embalado em caixas com 50 unidades. Sabe-se que as caixas vazias pesam 500 g, com desvio-padrão de 25 g. Admitindo-se uma distribuição normal dos pesos e independência entre as variáveis dos pesos do produto e da caixa, calcule a probabilidade de uma caixa cheia pesar mais de 1050 g.

$$\bar{X}_{\text{geral}} = 1000, V_{\text{geral}} = 50V_p + V_c, S_{\text{geral}} = \sqrt{V_{\text{geral}}} = 28.73 \text{ (R = 0.04093)}$$

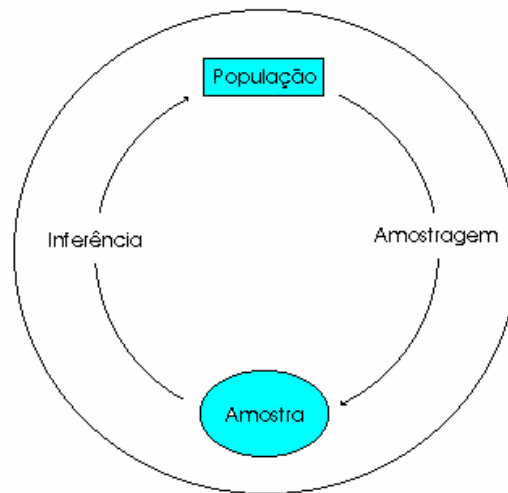
- 4) Em uma distribuição normal 28% dos elementos são superiores a 34 e 12% inferiores a 19. Encontrar a média e a variância da distribuição. (R =  $\bar{X} = 29.03$ ,  $S^2 = 73.44$ )
- 5) Suponha que a duração de vida de dois equipamentos  $E_1$  e  $E_2$  tenham respectivamente distribuições  $N(45 ; 9)$  e  $N(40 ; 36)$ . Se o equipamento tiver que ser usado por período de 45 horas, qual deles deve ser preferido? (R =  $E_1$ )
- 6) A precipitação pluviométrica média em certa cidade, no mês de dezembro, é de 8,9 cm. Admitindo a distribuição normal com desvio padrão de 2,5 cm, determinar a probabilidade de que, no mês de dezembro próximo, a precipitação seja (a) inferior a 1,6 cm, (b) superior a 5 cm mas não superior a 7,5 cm, (c) superior a 12 cm.
- 7) Em uma grande empresa, o departamento de manutenção tem instruções para substituir as lâmpadas antes que se queimem (não esperar que queimem para então substituí-las). Os registros indicam que a duração das lâmpadas tem distribuição  $N(900 ; 75)$  (horas). Quando devem ser substituídas as lâmpadas de modo que no máximo 10% delas queimem antes de serem trocadas? (R = 889 horas)
- 8) Os registros indicam que o tempo médio para se fazer um teste é aproximadamente  $N(80 ; 20)$  (min.). Determinar:
- a) a percentagem de candidatos que levam menos de 60 min ?
  - b) se o tempo concedido é de 1h, que percentagem não conseguirá terminar o teste ?
- 9) A profundidade dos poços artesianos em um determinado local é uma variável aleatória  $N(20 ; 3)$  (metros). Se  $X$  é a profundidade de determinado poço, determinar (a)  $P(X < 15)$ , (b)  $P(18 < X < 23)$ , (c)  $P(X > 25)$ .
- 10) Certa máquina de empacotar determinado produto oferece variações de peso com desvio padrão de 20 g. Em quanto deve ser regulado o peso médio do pacote para que apenas 10% tenham menos que 400 g? Calcule a probabilidade de um pacote sair com mais de 450 g. [R = a)  $\mu = 425.6$  b) 0.11123 ]]

# 7 Amostragem

## 7.1 Conceitos em Amostragem

**Inferência Estatística** - é o processo de obter informações sobre uma população a partir de resultados observados na Amostra.

**Amostragem:** É o processo de retirada de informações dos "n" elementos amostrais, na qual deve seguir um método adequado (tipos de amostragem).



## 7.2 Plano de Amostragem

- 1ª) Definir os Objetivos da Pesquisa
- 2ª) População a ser Amostrada
  - Parâmetros a ser Estimados (Objetivos)
- 3ª) Definição da Unidade Amostral
  - Seleção dos Elementos que farão parte da amostra
- 4ª) Forma de seleção dos elementos da população

- Tipo de Amostragem  $\left\{ \begin{array}{l} \text{Aleatoria Simples} \\ \text{Sistemática} \\ \text{Estratificada} \\ \text{por Conglomerados} \end{array} \right.$

5ª) Tamanho da Amostra

Ex.: Moradores de uma Cidade (população alvo)

Objetivo: Tipo de Residência  $\left\{ \begin{array}{l} \text{própria} \\ \text{alugada} \\ \text{emprestada} \end{array} \right. \left\{ \begin{array}{l} \text{um piso} \\ \text{dois pisos} \\ \text{tres ou mais pisos} \end{array} \right.$

Unidade Amostral: Domicílios (residências)

Elementos da População: Família por domicílio

Tipo de Amostragem:  $\left\{ \begin{array}{l} \text{aleatoria simples} \\ \text{sistemática} \\ \text{estratificada} \end{array} \right.$

## 7.3 Tipos de Amostragem

### 7.3.1 Amostragem Simples ou Ocasional

É o processo mais elementar e freqüentemente utilizado. Todos os elementos da população tem igual probabilidade de serem escolhidos. Para uma população finita o processo deve ser sem reposição. Todos os elementos da população devem ser numerados. Para realizar o sorteio dos elementos da população devemos usar a **Tabela de Números Aleatórios**.

### 7.3.2 Amostragem Sistemática

Trata-se de uma variação da Amostragem Aleatória Ocasional, conveniente quando a população está naturalmente ordenada, como fichas em um fichário, lista telefônica, etc.

Ex.:  $N = 5000$        $n = 50$ , então  $r = \frac{N}{n} = 10$ , (P.A. de razão 10)

Sorteia-se usando a **Tabela de Números Aleatórios** um número entre 1 e 10, ( $x=3$ ), o número sorteado refere-se ao 1º elemento da amostra, logo os elementos da amostra serão:

3    13    23    33    43    .....

Para determinar qualquer elemento da amostra podemos usar a fórmula do termo geral de uma P.A.

$$a_n = a_1 + (n - 1) \cdot r$$

### **7.3.3 Amostragem Estratificada**

É um processo de amostragem usado quando nos depararmos com populações heterogêneas, na qual pode-se distinguir subpopulações mais ou menos homogêneas, denominados estratos.

Após a determinação dos estratos, seleciona-se uma amostra aleatória de cada uma subpopulação (estrato).

As diversas subamostras retiradas das subpopulações devem ser proporcionais aos respectivos números de elementos dos estratos, e guardarem a proporcionalidade em relação a variabilidade de cada estrato, obtendo-se uma estratificação ótima.

Tipos de variáveis que podem ser usadas em estratificação: idade, classes sociais, sexo, profissão, salário, procedência, etc.

### **7.3.4 Amostragem por Conglomerados (ou Agrupamentos)**

Algumas populações não permitem, ou tornam-se extremamente difícil que se identifiquem seus elementos, mas podemos identificar subgrupos da população. Em tais casos, uma amostra aleatória simples desses subgrupos (conglomerados) podem ser escolhida, e uma contagem completa deve ser feita no conglomerado sorteado.

Agregados típicos são: quarteirões, famílias, organizações, agências, edifícios, etc.

## **7.4 Amostragem "COM" e "SEM" reposição**

Seja "N" o número de elementos de uma população, e seja "n" o número de elementos de uma amostra, então:

Se o processo de retirada dos elementos for COM reposição (pop. infinita ( $f \leq 5\%$ )), o número de amostras possíveis será:

$$n^{\text{º}} \text{ de amostras} = N^n$$

Se o processo de retirada de elementos for SEM reposição (pop. finita ( $f > 5\%$ )), o número de amostras possíveis será:

$$n^{\text{a}} \text{ de amostras} = C_{N,n} = \frac{N!}{n!(N-n)!}$$

Ex.: Supondo  $N = 8$  e  $n = 4$

com reposição:  $n^{\text{a}} \text{ de amostras} = N^n = 8^4 = 4096$

sem reposição:  $n^{\text{a}} \text{ de amostras} = C_{N,n} = \frac{N!}{n!(N-n)!} = C_{8,4} = \frac{8!}{4!4!} = 70$

Ex.: Processo de Amostragem Aleatória Simples  
(Distribuição Amostral das Médias)

- (com reposição)

$N = \{ 1, 2, 3, 4 \}$      $n = 2$      $n^{\text{a}} \text{ de amostras} = N^n = 4^2 = 16$

{1,1}	{1,2}	{1,3}	{1,4}
{2,1}	{2,2}	{2,3}	{2,4}
{3,1}	{3,2}	{3,3}	{3,4}
{4,1}	{4,2}	{4,3}	{4,4}

- (sem reposição)

$N = \{ 1, 2, 3, 4 \}$      $n = 2$      $n^{\text{a}} \text{ de amostras} = C_{4,2} = \frac{4!}{2!2!} = 6$

{1,2}	{1,3}	{1,4}
{2,3}	{2,4}	{3,4}

Para ilustrar melhor as estatísticas amostrais usaremos o processo com reposição.

{1,1} $\Rightarrow \bar{x} = 1,0$	{1,2} $\Rightarrow \bar{x} = 1,5$	{1,3} $\Rightarrow \bar{x} = 2,0$	{1,4} $\Rightarrow \bar{x} = 2,5$
{2,1} $\Rightarrow \bar{x} = 1,5$	{2,2} $\Rightarrow \bar{x} = 2,0$	{2,3} $\Rightarrow \bar{x} = 2,5$	{2,4} $\Rightarrow \bar{x} = 3,0$
{3,1} $\Rightarrow \bar{x} = 2,0$	{3,2} $\Rightarrow \bar{x} = 2,5$	{3,3} $\Rightarrow \bar{x} = 3,0$	{3,4} $\Rightarrow \bar{x} = 3,5$
{4,1} $\Rightarrow \bar{x} = 2,5$	{4,2} $\Rightarrow \bar{x} = 3,0$	{4,3} $\Rightarrow \bar{x} = 3,5$	{4,4} $\Rightarrow \bar{x} = 4,0$

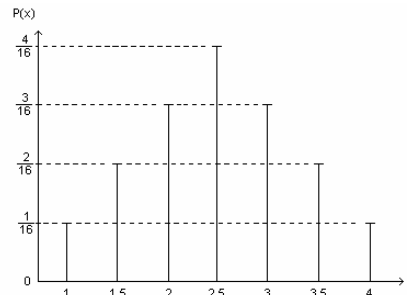


## 7.5 Representações de uma Distribuição Amostral

- Tabela

$\bar{x}_i$	$P(\bar{X} = \bar{x}_i)$
1,0	1/16
1,5	2/16
2,0	3/16
2,5	4/16
3,0	3/16
3,5	2/16
4,0	1/16
$\Sigma$	16/16

- Gráfico



## 7.6 Estatísticas Amostrais

- Esperança Matemática

$$\mu(\bar{x}) = E(\bar{x}) = \sum_{i=1}^n \bar{x}_i P(\bar{X} = \bar{x}_i) = \frac{40}{16} = 2,5$$

-Variância

$$\text{ou } \text{VAR}(\bar{x}) = E(\bar{x}^2) - [E(\bar{x})]^2$$

$$\text{onde } E(\bar{x}^2) = \sum_{i=1}^n \bar{x}_i^2 P(\bar{X} = \bar{x}_i)$$

## 7.7 TAMANHO DA AMOSTRA

### 7.7.1 Introdução

Os pesquisadores de todo o mundo, na realização de pesquisas científicas, em qualquer setor da atividade humana, utilizam as técnicas de amostragem no planejamento de seus trabalhos, não só pela impraticabilidade de poderem observar, numericamente, em sua totalidade determinada população em estudo, como devido ao aspecto econômico dessas investigações, conduzidos com um menor custo operacional, dentro de um menor tempo, além de possibilitar maior precisão nos respectivos resultados, ao contrário, do que ocorre com os trabalhos realizados pelo processo censitário (COCHRAN, 1965; CRUZ, 1978).

A técnica da amostragem, a despeito de sua larga utilização, ainda necessita de alguma didática mais adequada aos pesquisadores iniciantes.

Na teoria da amostragem, são consideradas duas dimensões:

1ª) Dimensionamento da Amostra;

2ª) Composição da Amostra.

### 7.7.2 Procedimentos para determinar o tamanho da amostra

1ª) Analisar o questionário, ou roteiro da entrevista e escolher uma variável que julgue mais importante para o estudo. Se possível mais do que uma;

2ª) Verificar o nível de mensuração da variável: nominal, ordinal ou intervalar;

3ª) Considerar o tamanho da população: infinita ou finita

4ª) Se a variável escolhida for:

- **intervalar e a população considerada infinita, você poderá determinar o tamanho da amostra pela fórmula:**

$$n = \left( \frac{Z \cdot \sigma}{d} \right)^2$$

onde:  $Z$  = abscissa da curva normal padrão, fixado um nível de confiança  $(1 - \alpha)$

$$Z = 1,65 \rightarrow (1 - \alpha) = 90\%$$

$$Z = 1,96 \rightarrow (1 - \alpha) = 95\%$$

$$Z = 2,0 \rightarrow (1 - \alpha) = 95.5\%$$

$$Z = 2,57 \rightarrow (1 - \alpha) = 99\%$$

Geralmente usa-se  $Z = 2$

$\sigma$  = desvio padrão da população, expresso na unidade variável, onde poderá ser determinado por:

- Especificações Técnicas
- Resgatar o valor de estudos semelhantes
- Fazer conjecturas sobre possíveis valores

$d$  = erro amostral, expresso na unidade da variável. O erro amostral é a máxima diferença que o investigador admite suportar entre  $\mu$  e  $\bar{x}$ , isto é:  $|\mu - \bar{x}| < d$ .

- **intervalar e a população considerada finita, você poderá determinar o tamanho da amostra pela fórmula:**

$$n = \frac{Z^2 \cdot \sigma^2 \cdot N}{d^2(N - 1) + Z^2 \cdot \sigma^2}$$

onde:  $Z$  = abscissa da normal padrão

$\sigma^2$  = variância populacional

$N$  = tamanho da população

$d$  = erro amostral

- **nominal ou ordinal, e a população considerada infinita, você poderá determinar o tamanho da amostra pela fórmula:**

$$n = \frac{Z^2 \cdot \hat{p} \cdot \hat{q}}{d^2}$$

onde:  $Z$  = abscissa da normal padrão

$\hat{p}$  = estimativa da verdadeira proporção de um dos níveis da variável escolhida. Por exemplo, se a variável escolhida for parte da empresa,  $\hat{p}$  poderá ser a estimativa da

verdadeira proporção de grandes empresas do setor que está sendo estudado.  $\hat{p}$  será expresso em decimais ( $\hat{p} = 30\% \rightarrow \hat{p} = 0.30$ ).

$$\hat{q} = 1 - \hat{p}$$

$d$  = erro amostral, expresso em decimais. O erro amostral neste caso será a máxima diferença que o investigador admite suportar entre  $\pi$  e  $\hat{p}$ , isto é:  $|\pi - \hat{p}| < d$ , em que  $\pi$  é a verdadeira proporção (frequência relativa do evento a ser calculado a partir da amostra).

- **nominal ou ordinal, e a população considerada finita, você poderá determinar o tamanho da amostra pela fórmula:**

$$n = \frac{Z^2 \cdot \hat{p} \cdot \hat{q} \cdot N}{d^2(N-1) + Z^2 \cdot \hat{p} \cdot \hat{q}}$$

onde:  $Z$  = abscissa da normal padrão

$N$  = tamanho da população


$\hat{p}$  = estimativa da proporção

$$\hat{q} = 1 - \hat{p}$$

$d$  = erro amostral

Estas fórmulas são básicas para qualquer tipo de composição da amostra; todavia, existem fórmulas específicas segundo o critério de composição da amostra.

- Se o investigador escolher mais de uma variável, poderá acontecer de ter que aplicar mais de uma fórmula, assim deverá optar pelo maior valor de "n".

 Quando não tivermos condições de prever o possível valor para  $\hat{p}$ , admita  $\hat{p} = 0.50$ , pois, dessa forma, você terá o maior tamanho da amostra, admitindo-se constantes os demais elementos.

## 7.8 Distribuições amostrais de probabilidade

### 7.8.1 Distribuição amostral das médias

Se a variável aleatória "x" segue uma distribuição normal:

$$\bar{x} \sim N(\mu(\bar{x}); \sigma^2(\bar{x})), \text{ onde } z = \frac{\bar{x} - \mu(\bar{x})}{\sigma(\bar{x})}$$

$\mu(\bar{x}) = \mu$  (a média amostral é igual a média populacional) e  $\sigma(\bar{x}) = \frac{\sigma(x)}{\sqrt{n}}$  (Desvio Padrão Amostral)

### 7.8.1.1 Caso COM reposição (pop. infinita)

$$\bar{x} \sim N\left[\mu(\bar{x}); \frac{\sigma^2(x)}{n}\right]$$

### 7.8.1.2 Caso SEM reposição (pop. finita)

Quando a amostra for  $> 5\%$  da população  $\left(\frac{n}{N}\right)$  devemos usar um fator de correção.

$$\bar{x} \sim N\left[\mu(\bar{x}); \frac{\sigma^2(x)}{n} \frac{N-n}{N-1}\right], \text{ onde } \frac{N-n}{N-1} \text{ é o fator de correção}$$

Ex<sub>1</sub>.: Uma população muito grande tem média 20,0 e desvio padrão 1,4 . Extraí-se uma amostra de 49 observações. Responda:

- Qual a média da distribuição amostral ?
- Qual o desvio padrão da distribuição amostral ?
- Qual a porcentagem das possíveis médias que diferiram por mais de 0,2 da média populacional ?

Ex<sub>2</sub>.: Um processo de encher garrafas de coca-cola dá em média 10% mal cheias com desvio padrão de 30%. Extraída uma amostra de 225 garrafas de uma sequência de produção de 625, qual a probabilidade amostral das garrafas mal cheias estar entre 9% e 12%.

O exemplo n 2 pode ser resolvido usando a distribuição amostral das proporções, onde  $p$  = proporção populacional,  $\bar{p}$  = média da distribuição amostral das proporções. Logo temos:

### 7.8.2 Distribuição amostral das proporções

$$p = \bar{p} \quad \text{e} \quad \bar{\sigma}_p = \sqrt{\frac{p(1-p)}{n}} \cdot \sqrt{\frac{N-n}{N-1}},$$

onde  $\sqrt{\frac{N-n}{N-1}}$  é usado para população finita.

Ex<sub>1</sub>: Uma máquina de recobrir cerejas com chocolate é regulada para produzir um revestimento de (3% em relação ao volume da cereja). Se o processo segue uma distribuição normal, qual a probabilidade de extrair uma amostra de 25 cerejas de um lote de 169 e encontrar uma média amostral superior a 3,4%. R = 0,44828.

### 7.9 Exercícios

- 1) Uma fabrica de baterias alega que eu artigo de primeira categoria tem uma vida média de 50 meses, e desvio padrão de 4 meses.
  - a) Que porcentagem de uma amostra de 36 observações acusaria vida média no intervalo de um mês em torno da média ?
  - b) Qual será a resposta para uma amostra de 64 observações?
  - c) Qual seria o percentual das médias amostrais inferior a 49,8 meses com n =100?
- 2) Um varejista compra copos diretamente da fábrica em grandes lotes. Os copos são embrulhados individualmente. Periodicamente o varejista inspeciona os lotes para determinar a proporção dos quebrados ou lascados. Se um grande lote contém 10% de quebrados (lascados) qual a probabilidade do varejista obter numa a mostra de 100 copos 17% ou mais defeituosos?
- 3) Deve-se extrair uma amostra de 36 observações de uma máquina de cunhar moedas comemorativas. A espessura média das moedas é de 0,2 cm, com desvio padrão de 0,01 cm.
  - a) Que porcentagem de médias amostrais estará no intervalo  $\pm 0,004$  em torno da média? R = 0.98316
  - b) Qual a probabilidade de obter uma média amostral que se afaste por mais de 0,005 cm da média do processo ? R = 0.00164
- 4) Suponha que uma pesquisa recente tenha revelado que 60% de uma população de adultos do sexo masculino consistam de não-fumantes. Tomada uma amostra de 10 pessoas de uma população muito grande, que porcentagem esperamos nos intervalos abaixo:
  - a) de 50% a 65%
  - b) maior que 53%
  - c) de 65% a 80%
- 5) Se a vida média de operação de um "flash" é 24 horas, com distribuição normal e desvio padrão de 3 horas, qual é a probabilidade de uma amostra de 10 "flashes" retirados de uma população de 500 "flashes" apresentar vida média que difira por mais de 30 min. da média. R = 0.60306

## 8 Estimação de Parâmetros

É um processo de indução, na qual usamos dados extraídos de uma amostra para produzir inferência sobre a população. Esta inferência só será válida se a amostra for significativa.

### - Tipos de Estimações de Parâmetros

- i) Estimação Pontual
- ii) Estimação Intervalar

#### 8.1 Estimação Pontual

É usada quando a partir da amostra procura-se obter um único valor de certo parâmetro populacional, ou seja, obter estimativas a partir dos valores amostrais.

##### a) Estatísticas

Seja  $(X_1, X_2, \dots, X_n)$  uma amostra aleatória e  $(x_1, x_2, \dots, x_n)$  os valores tomados pela amostra; então  $y = H(x_1, x_2, \dots, x_n)$  é uma estatística.

Principais estatísticas:

- Média Amostral
- Proporção Amostral
- Variância Amostral

#### 8.2 Estimação Intervalar

Uma outra maneira de se calcular um estimativa de um parâmetro desconhecido, é construir um intervalo de confiança para esse parâmetro com uma probabilidade de  $1 - \alpha$  (nível de confiança) de que o intervalo contenha o verdadeiro parâmetro. Dessa maneira  $\alpha$  será o nível de significância, isto é, o erro que se estará cometendo ao afirmar que o parâmetro está entre o limite inferior e o superior calculado.

##### 8.2.1 Intervalo de confiança para a média ( $m$ ) com a variância ( $\sigma^2$ ) conhecida.

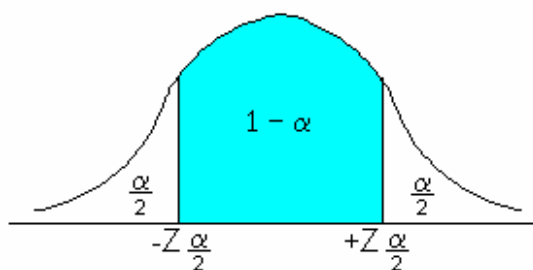
$(n > 30 \rightarrow Z)$

Seja  $X \sim N(\mu, \sigma^2)$

Como já vimos anteriormente,  $\bar{x}$  (média amostral) tem distribuição normal de média  $\mu$  e desvio padrão  $\frac{\sigma}{\sqrt{n}}$ , ou seja:

$$X \sim N\left(\mu; \frac{\sigma^2}{n}\right)$$

Portanto,  $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  tem distribuição  $N(0,1)$



Então,

$$P(-z_{\alpha/2} \leq z \leq +z_{\alpha/2}) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq +z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} - \bar{X} \leq \mu \leq +z_{\alpha/2} \frac{\sigma}{\sqrt{n}} - \bar{X}\right) = 1 - \alpha$$

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad (\text{Pop. Infinita})$$

Para caso de populações finitas usa-se a seguinte fórmula:

$$P\left(\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}\right) = 1 - \alpha \quad (\text{Pop. Finita})$$

Obs.: Os níveis de confiança mais usados são:

$$1 - \alpha = 90\% \Rightarrow z_{\alpha/2} = \pm 1,64$$

$$1 - \alpha = 95\% \Rightarrow z_{\alpha/2} = \pm 1,96$$

$$1 - \alpha = 99\% \Rightarrow z_{\alpha/2} = \pm 2,58$$

$$1 - \alpha = 85\% \Rightarrow z_{\alpha/2} = \underline{\hspace{2cm}}$$



Ex.: Seja X a duração da vida de uma peça de equipamento tal que  $\sigma = 5$  horas. Admita que 100 peças foram ensaiadas fornecendo uma duração de vida média de 500 horas e que se deseja obter um intervalo de 95% para a verdadeira média populacional.  $R = P(499,02 \leq \mu \leq 500,98) = 95\%$ .

Obs.: Podemos dizer que 95% das vezes, o intervalo acima contém a verdadeira média populacional. Isto não é o mesmo que afirmar que 95% é a probabilidade do parâmetro  $\mu$  cair dentro do intervalo, o que constituirá um erro, pois  $\mu$  é um parâmetro (número) e ele está ou não no intervalo.

### 8.2.2 Intervalo de confiança para a média ( $m$ ) com a variância ( $\sigma^2$ ) desconhecida

( $n \leq 30$ )

Neste caso precisa-se calcular a estimativa S (desvio padrão amostral) a partir dos dados, lembrando que:

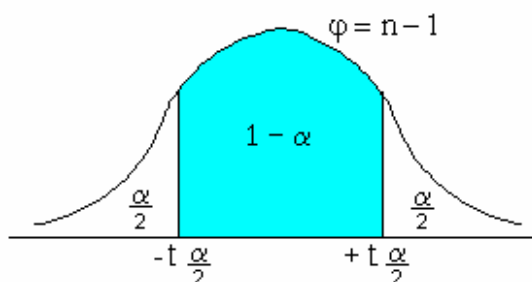
$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad \text{onde } n-1 = \text{graus de liberdade}$$

$$\bar{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$$

Portanto,  $t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$  tem distribuição N(0,1)

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \cdot \frac{\sigma}{S} = \frac{z}{S/\sigma} = \frac{N(0,1)}{S/\sigma}$$

Esta distribuição é conhecida como distribuição "t" de Student, no caso com ( $\varphi = n - 1$ ) graus de liberdade



O gráfico da função densidade da variável "t" é simétrico e tem a forma da normal, porém menos "achatada" sua média vale 0 e a variância  $\frac{\varphi}{\varphi - 2}$  em que  $\varphi$  é o grau de liberdade ( $\varphi > 2$ )

$$t_{\varphi, \alpha/2} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Então,

$$P(-t_{\varphi, \alpha/2} \leq t \leq +t_{\varphi, \alpha/2}) = 1 - \alpha$$

$$P\left(\bar{X} - t_{\varphi, \alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\varphi, \alpha/2} \frac{S}{\sqrt{n}}\right) = 1 - \alpha \quad (\text{Pop. Infinita})$$

Para caso de populações finitas usa-se a seguinte fórmula:

$$P\left(\bar{X} - t_{\varphi, \alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{X} + t_{\varphi, \alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}\right) = 1 - \alpha \quad (\text{Pop. Finita})$$

Ex.: A seguinte amostra: 9, 8, 12, 7, 9, 6, 11, 6, 10, 9 foi extraída de uma população aproximadamente normal. Construir um intervalo de confiança para  $\mu$  com um nível de 95%.

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = 8,7 \quad S^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1} \cong 4 \quad S \cong 2$$

$$\varphi = 10 - 1 = 9 \quad \alpha = 5\%$$

$$t_{\varphi, \alpha/2} = t_{9, 2,5\%} = \pm 2,262$$

$$R = P(7,27 \leq \mu \leq 10,13) = 95\%$$

Obs.: Quando  $n > 30$  e  $\sigma$  for desconhecido poderemos usar S como uma boa estimativa de  $\sigma$ . Esta estimação será melhor quanto maior for o tamanho da amostra.

### 8.2.3 Intervalo de Confiança para Proporções

Sendo  $\hat{p}$  o estimador de  $\pi$ , onde  $\hat{p}$  segue uma distribuição normal, logo:

$$\hat{p} \sim N\left(\hat{p}; \frac{\hat{p} \cdot \hat{q}}{n}\right) \quad (\text{pop. infinita})$$

$$\hat{p} \sim N\left[\hat{p}; \frac{\hat{p} \cdot \hat{q}}{n} \left(\frac{N-n}{N-1}\right)\right] \quad (\text{pop. finita})$$

$$\text{logo } Z = \frac{\hat{p} - \pi}{\sigma_{\hat{p}}} \quad \text{onde } \begin{cases} \sigma_{\hat{p}} = \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} \\ \hat{p} = \frac{X}{n} = \frac{\text{característica}}{\text{número de elementos da amostra}} \end{cases}$$

$$P(\hat{p} - Z_{\alpha/2} \sigma_{\hat{p}} \leq \pi \leq \hat{p} + Z_{\alpha/2} \sigma_{\hat{p}}) = 1 - \alpha \quad (\text{Pop. Infinita})$$

Para caso de populações finitas usa-se a seguinte fórmula:

$$P\left(\hat{p} - Z_{\alpha/2} \sigma_{\hat{p}} \sqrt{\frac{N-n}{N-1}} \leq \pi \leq \hat{p} + Z_{\alpha/2} \sigma_{\hat{p}} \sqrt{\frac{N-n}{N-1}}\right) = 1 - \alpha \quad (\text{Pop. Finita})$$

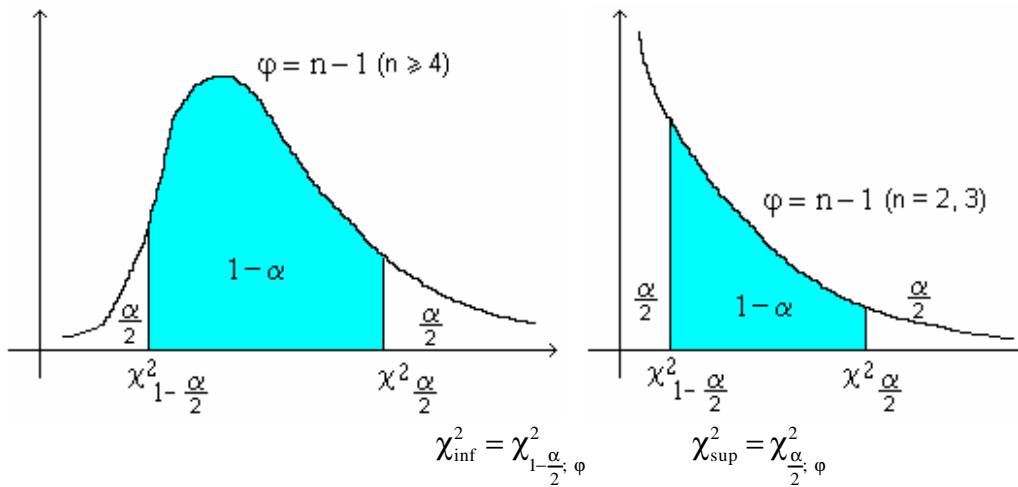
Ex.: Uma centena de componentes eletrônicos foram ensaiados e 93 deles funcionaram mais que 500 horas. Determine um intervalo de confiança de 95% para a verdadeira proporção populacional sabendo que os mesmos foram retirados de uma população de 1000 componentes.

#### 8.2.4 Intervalo de Confiança para Variância

Como o estimador de  $\sigma^2$  é  $S^2$  pode-se considerar que  $\frac{(n-1)S^2}{\sigma^2}$  tem distribuição Qui - quadrado, ou seja:

$$X_{n-1}^2 \sim Z \frac{S^2}{\sigma^2},$$

logo o intervalo será:



Assim temos:

$$P\left[\frac{(n-1)S^2}{\chi_{sup}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{inf}^2}\right] = 1 - \alpha$$

Ex.: A seguinte amostra: 9, 8, 12, 7, 9, 6, 11, 6, 10, 9 foi extraída de uma população aproximadamente normal. Construir um intervalo de confiança para  $\sigma^2$  com um nível de 95%.

### 8.2.5 Intervalo de Confiança para a diferença Entre duas Médias:

Usualmente comparamos as médias de duas populações formando sua diferença:

$$\mu_1 - \mu_2$$

Uma estimativa pontual desta diferença correspondente:

$$\bar{X}_1 - \bar{X}_2$$

#### a) Variâncias Conhecidas

$$m_1 - m_2 = (\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \cdot (\text{Erro Padrão})$$

Erro Padrão?

$$\begin{aligned} \text{VAR}(\bar{X}_1 - \bar{X}_2) &= (+1)^2 \text{VAR} \bar{X}_1 + (-1)^2 \text{VAR} \bar{X}_2 \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{aligned}$$

$$\text{logo o erro padrão} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$P\left[(\bar{X}_1 - \bar{X}_2) - Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] = 1 - \alpha$$

Obs.: se  $\sigma_1$  e  $\sigma_2$  são conhecidas e tem um valor em comum, logo:

$$\text{Erro Padrão: } \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Ex.: Seja duas classes muito grande com desvios padrões  $\sigma_1 = 1,21$  e  $\sigma_2 = 2,13$ . Extraída uma amostra de 25 alunos da classe 1 obteve-se uma nota média de 7,8, e da classe 2 foi extraída uma amostra de 20 alunos obteve-se uma nota média de 6,0. Construir um intervalo de 95% de confiança para a verdadeira diferença das médias populacionais. R = (LI=0,753; LS=2,847)

### b) Variâncias Desconhecidas

Em geral conhecemos duas variâncias populacionais ( $\sigma_1^2$  e  $\sigma_2^2$ ). Se as mesmas são desconhecidas o melhor que podemos fazer é estimá-las por meio de variâncias amostrais  $S_1^2$  e  $S_2^2$ .

Como as amostras serão pequenas, introduziremos uma fonte de erro compensada pela distribuição "t":

$$P\left[(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right] \text{ onde } \phi = n_1 + n_2 - 2$$

Obs.: Se as variâncias populacionais são desconhecidas mas as estimativas são iguais, poderemos usar para o Erro Padrão o seguinte critério:

$$\text{Erro Padrão: } S_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{onde } S_c \text{ é o desvio padrão conjunto}$$

$$S_c = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Ex<sub>1</sub>.: De uma turma (1) foi extraída uma amostra de 6 alunos com as seguintes alturas: 150, 152, 153, 160, 161, 163. De uma segunda turma foi extraída uma amostra de 8 alunos com as seguintes alturas: 165, 166, 167, 172, 178, 180, 182, 190. Contruir um intervalo de 95% de confiança para a verdadeira diferença entre as médias populacionais.

Ex<sub>2</sub>.: De uma máquina foi extraída uma amostra de 8 peças, com os seguintes diâmetros: 54, 56, 58, 60, 60, 62, 63, 65. De uma segunda máquina foi extraída uma amostra de 10 peças, com os seguintes diâmetros: 75, 75, 76, 77, 78, 78, 79, 80, 80, 82. Construir um intervalo de 99% de confiança para a diferença entre as médias populacionais, supondo que as máquinas foram construídas pelo mesmo fabricante.

### 8.2.6 Intervalo de Confiança para a Diferença entre duas Proporções

$$P \left[ (\hat{p}_1 - \hat{p}_2) - Z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \leq \pi_1 - \pi_2 \leq (\hat{p}_1 - \hat{p}_2) + Z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right] = 1 - \alpha$$

Ex.: Em uma pesquisa realizada pelo Instituto Gallup constatou que 500 estudantes entrevistados com menos de 18 anos, 50% acreditam na possibilidade de se verificar uma modificação na América, e que dos 100 estudantes com mais de 24 anos, 69% acreditam nessa modificação. Construir um intervalo de confiança para a diferença entre as proporções destas subpopulações usando  $\alpha=5\%$ . R=(LI=0,0893, LS=0,2905)

### 8.3 Exercícios

- 1) Ao se realizar uma contagem de eritrócitos em 144 mulheres encontrou-se em média 5,35 milhões e desvio padrão 0,4413 milhões de glóbulos vermelhos. Determine os limites de confiança de 99% para a média populacional.
- 2) Um conjunto de 12 animais de experiência receberam uma dieta especial durante 3 semanas e produziram os seguintes aumentos de peso (g): 30, 22, 32, 26, 24, 40, 34, 36, 32, 33, 28 e 30. Determine um intervalo de 90% de confiança para a média.  
R  $\Rightarrow \bar{X}=30.58, S = 5.09, LI = 27.942, LS = 33.218$
- 3) Construa um intervalo de 95 % de confiança para um dos seguintes casos:

	Média Amostral	$\sigma$	Tamanho da Amostra
a)	16,0	2,0	16
b)	37,5	3,0	36
c)	2,1	0,5	25
d)	0,6	0,1	100

- 4) Numa tentativa de melhorar o esquema de atendimento, um médico procurou estimar o tempo médio que gasta com cada paciente. Uma amostra aleatória de 49 pacientes, colhida num período de 3 semanas, acusou uma média de 30 min., com desvio padrão de 7 min. Construa um intervalo de 95% de confiança para o verdadeiro tempo médio de consulta.

- 5) Solicitou-se a 100 estudantes de um colégio que anotasse suas despesas com alimentação e bebidas no período de uma semana. Há 500 estudantes no colégio. O resultado foi uma despesa de \$40,00 com um desvio padrão de \$10,00. Construa um intervalo de 95% de confiança para a verdadeira média.
- 6) Uma amostra aleatória de 100 fregueses da parte da manhã de um supermercado revelou que apenas 10 não incluem leite em suas compras.
- a) qual seria a estimativa percentual dos fregueses que compram leite pela parte da manhã. ( $\alpha = 5\%$ ).  $R\hat{P}$   $LI = 84.12\%$ ,  $LS = 95.88\%$
- b) construir um intervalo de 90% de confiança para a verdadeira proporção dos fregueses que não compram leite pela manhã.  $R\hat{P}$   $LI = 5.08\%$ ,  $LS = 14.92\%$
- 7) Uma amostra aleatória de 40 homens trabalhando num grande projeto de construção revelou que 6 não estavam usando capacetes protetores. Construa um intervalo de 98% de confiança para a verdadeira proporção dos que não estão usando capacetes neste projeto.  $R\hat{P}$   $S_{\hat{p}} = 0.056$ ,  $LI = 0.02$ ,  $LS = 0.28$
- 8) De 48 pessoas escolhidas aleatoriamente de uma longa fila de espera de um cinema, 25% acharam que o filme principal continha demasiada violência.
- a) qual deveria ser o tamanho da fila, a partir do qual se pudesse desprezar o fator de correção finita;
- b) construa um intervalo de 98% de confiança para a verdadeira proporção, se há 100 pessoas na fila;
- c) construa um intervalo de 98% de confiança para a verdadeira proporção, se há 500 pessoas na fila.
- 9) Em uma fábrica, colhida uma amostra ( $n = 30$ ) de certa peça, obtiveram-se as seguintes medidas para os diâmetros:

10	11	11	11	12	12	12	12	13	13
13	13	13	13	13	13	13	13	13	13
14	14	14	14	14	15	15	15	16	16

- a) estimar a média e a variância;  $R\hat{P}$   $13,13$  e  $2,05$
- b) construir um intervalo de confiança para a média, sendo  $\alpha = 5\%$ .  
( $LI = 12.536$  e  $LS = 13.664$ )
- c) construir um intervalo de 95 % de confiança para a média, supondo que a amostra tenha sido retirada de uma população de 100 peças, sendo  $\alpha = 5\%$ . ( $LI = 12.581$  e  $LS = 13.579$ )
- d) Construir um intervalo de confiança para a variância populacional, sendo  $\alpha = 5\%$ . ( $LI = 1.3003$  e  $LS = 3.704$ )
- 10) Supondo populações normais, contruir um intervalo de confiança para a média e para a variância ao nível de significância de 90% para as amostras.

a)	2	3	4	5	5	6	6	7	8	8	9	$n = 11$	
b)	12	12	15	15	16	16	17	18	20	22	22	23	$n = 12$
c)	25	25	27	28	30	33	34	35	36				$n = 9$

- 11) Sendo  $X$  uma população tal que  $X \sim N(\mu; \sigma^2)$  em que  $\mu$  e  $\sigma^2$  são desconhecidos. Uma amostra de tamanho 15 forneceu os seguintes valores  $\sum X_i = 8,7$ , e  $\sum X_i^2 = 27,3$ . Determinar um intervalo de confiança de 95% para  $\mu$  e  $\sigma^2$ , supondo:

$$\bar{X} = \frac{\sum X_i}{n} \quad e \quad S^2 = \frac{\sum X_i^2 - \left(\frac{\sum X_i}{n}\right)^2}{n-1}$$

- 12) Dados os seguintes conjuntos de medias, determinar um intervalo de 99% de confiança para a variância populacional e para a média populacional.

0.0105; 0,0193; 0,0152; 0,0229; 0,0244; 0,0190; 0,0208; 0,0253; 0,0276

$$R \text{ P } \bar{X} = 0.0206, LI = 0.01467, LS = 0.026653; \\ S = 0.0053, LI = 0.00001, LS = 0.000167$$

- 13) O tempo de reação a uma injeção intravenosa é em média de 2.1 min., com desvio padrão de 0.1 min., para grupos de 20 pacientes. Construa um intervalo de confiança de 90 % para o tempo médio para toda a população dos pacientes submetidos ao tratamento.

- 14) Uma firma esta convertendo as máquinas que aluga para uma versão mais moderna. Até agora foram convertidas 40 máquinas. O tempo médio de conversão foi de 24 horas com desvio padrão de 3 horas. a) Determine um intervalo de confiança de 99 % para o tempo médio de conversão. b) Para uma amostra de 60 máquinas, como ficaria o intervalo de confiança de 99 % para o tempo médio de conversão

- 15) Seis dentre 48 terminais telefônicos dão respostas de ocupado. Uma firma possui 800 terminais. Construir um intervalo de confiança de 95 % para a proporção dos terminas da firma que apresentam sinal de ocupado.  $R \text{ P } LI = 3.4\%, LS = 21.6\%$

- 16) Uma amostra de 50 bicicletas de um estoque de 400 bicicletas, acusou 7 com pneus vazios. a) Estime o número de bicicletas com pneus vazios: b) Construa um intervalo de 99 % confiança para a população das bicicletas com pneus vazios.

$$R \text{ P } a) 56; b) LI = 0.021 LS = 0.258$$

- 17) A média salarial semanal para uma amostra de  $n = 30$  empregados em uma grande firma é  $\bar{X} = 180,00$  com desvio padrão  $S = 14,00$ . Construa um intervalo de confiança de 99 % para a média salarial dos funcionários.

- 18) Uma empresa de pesquisa de mercado faz contato com uma amostra de 100 homens em uma grande comunidade e verifica que uma proporção de 0.40 na amostra prefere lâminas de barbear fabricadas por seu cliente em vez de qualquer outra marca. Determinar o intervalo de confiança de 95 % para a proporção de todos os homens na comunidade que preferem a lâmina do cliente.



- 19) Uma pequena fábrica comprou um lote de 200 pequenas peças eletrônicas de um saldo de estoque de uma grande firma. Para uma amostra aleatória de 50 peças, constatou-se que 5 eram defeituosas. Estimar a proporção de todas as peças que são defeituosas no carregamento utilizando um intervalo de confiança de 99 %.
- 20) Duas amostras de plantas foram cultivadas com dois fertilizantes diferentes. A primeira amostra oriunda de 200 sementes, acusou altura média de 10,9 cm e desvio padrão 2,0 cm. A segunda amostra, de 100 sementes, acusou uma altura média de 10,5 cm com desvio padrão de 5,0 cm. Construir um intervalo de confiança entre as alturas médias das populações ao nível de 95% de confiança.
- 21) Uma amostra aleatória de 120 trabalhadores de uma grande fábrica leva em média 22,0 min. para executar determinado serviço, com  $S^2$  de 4 min<sup>2</sup>. Em uma segunda fábrica, para executar a mesma tarefa, uma amostra aleatória de 120 operários, gasta em média 19,0 min com  $S^2$  de 10 min<sup>2</sup>. Construir um intervalo de 99% de confiança entre as médias das populações.
- 22) Extraíram-se amostras independentes de adultos brancos e pretos, que acusaram os seguintes tempos (em anos) de escolaridade.

Branco	8	18	10	10	14			
Preto	9	12	5	10	14	12	6	

Construir um intervalo de 95% de confiança para:

- a) a média da população branca, e a média da população preta;  
 b) a diferença entre as médias entre brancos e pretos
- 23) Em uma amostra aleatória de cinco pessoas, foi medida a capacidade torácica antes e após determinado tratamento, obtendo-se os dados a seguir. Construir um intervalo de 95% de confiança para a diferença entre as médias antes e depois do tratamento.

Pessoa	Capacidade Torácica	
	Antes (X)	Após (Y)
A	2750	2850
B	2360	2380
C	2950	2930
D	2830	2860
E	2250	2320

- 24) Extraída duas amostras de professores homens e mulheres, obteve-se os seguintes resultados quantos aos salários em milhares de dólares: Construir um intervalo de 95% de confiança para:

Homens	Mulheres
$n_1 = 25$	$n_2 = 5$
$\bar{X}_1 = 16,0$	$\bar{X}_2 = 11,0$
$S_1^2 = 16$	$S_2^2 = 10$

- a) a média da diferença entre os salários;  
 b) a média do salário dos homens;  
 c) a média do salário das mulheres.

- 25) Em uma pesquisa efetuada em com 1650 americanos foram consultados sobre o seguinte tema: "A mulher grávida pode procurar um médico e interromper a gravidez, a qualquer momento durante os três primeiros meses. É a favor ou contra esta decisão?" Uma semana mais tarde foram consultados 1650 americanos foram consultados sobre o mesmo assunto , exceto que a pergunta "a favor do aborto, ao invés de "interromper a gravidez". As respostas foram as seguintes:

Pergunta	Resposta		
	A favor	Contra	Sem opinião
"Interromper a gravidez"	46%	39%	15%
"A favor do aborto"	41%	49%	10%

- a) Seja  $\pi_1$  a proporção dos votantes a favor do 1º caso (interromper) e  $\pi_2$  a proporção dos votantes a favor do 2º caso (aborto). Construir um intervalo de 95% de confiança para a diferença  $\pi_1 - \pi_2$ .  $R = LI = 0.01628$ ,  $LS = 0.08379$

- b) Repetir a (a) para os votantes que não tiveram opinião.  $R = LI = 0.0275$ ,  $LS = 0.0725$

- 26) Numa pesquisa sobre intenção do comprador brasileiro. 30 famílias de uma amostra aleatória de 150 declararam ser uma intenção comprar um carro novo dentro de um ano. Uma outra amostra de 160 famílias 25 declararam a mesma intenção. Construir um intervalo de 99% de confiança para as diferenças entre as proporções.

$$R = LI = 0.0727, LS = 0.1527$$

## 9 Teste de Hipóteses

Trata-se de uma técnica para se fazer inferência estatística. Ou seja, a partir de um teste de hipóteses, realizado com os dados amostrais, pode-se inferir sobre a população.

No caso da inferência através de Intervalo de confiança, busca-se "cercar" o parâmetro populacional desconhecido. Aqui formula-se uma hipótese quanto ao valor do parâmetro populacional, e pelos elementos amostrais faz-se um teste que indicará a ACEITAÇÃO ou REJEIÇÃO da hipótese formulada.

### 9.1 Principais Conceitos

#### 9.1.1 Hipótese Estatística

Trata-se de uma suposição quanto ao valor de um parâmetro populacional, ou quanto à natureza da distribuição de uma variável populacional.

São exemplos de hipóteses estatísticas:

- a) A altura média da população brasileira é 1,65 m, isto é:  $H: \mu = 1,65$  m;
- b) A variância populacional dos salários vale \$ 500<sup>2</sup>, isto é,  $H: \sigma^2 = 500^2$ ;
- c) A proporção de paulistas fumantes é 25%, ou seja,  $H: p = 0.25$
- d) A distribuição dos pesos dos alunos da nossa faculdade é normal.

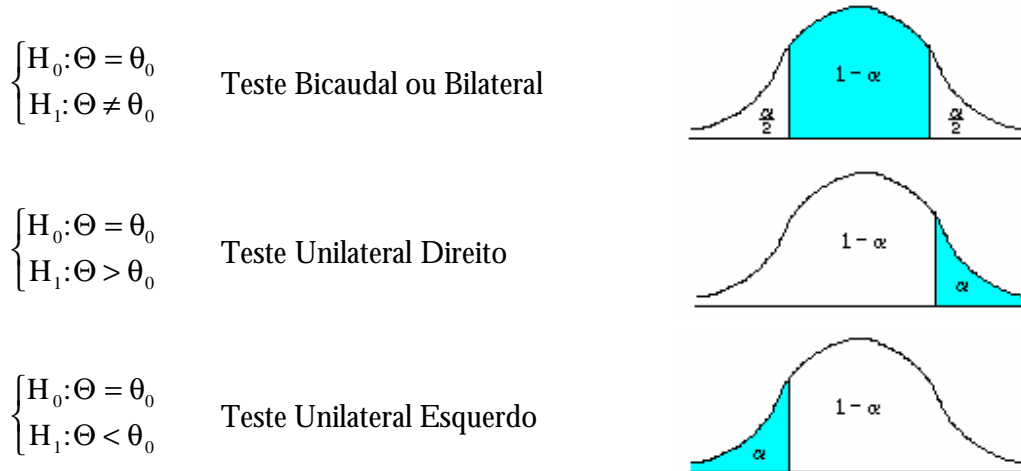
#### 9.1.2 Teste de Hipótese

É uma regra de decisão para aceitar ou rejeitar uma hipóteses estatística com base nos elementos amostrais.

#### 9.1.3 Tipos de Hipóteses

Designa-se por  $H_0$ , chamada hipótese nula, a hipóteses estatística a ser testada, e por  $H_1$  a hipótese alternativa. A hipótese nula expressa uma igualdade, enquanto que a hipótese alternativa é dada por uma desigualdade ( $\neq$ ,  $<$ ,  $>$ ).

Exemplos:



### 9.1.4 Tipos de erros

Há dois tipos de erro ao testar uma hipótese estatística. Pode-se rejeitar uma hipóteses quando ela é, de fato verdadeira, ou aceitar uma hipóteses quando ela é, de fato, falsa. A rejeição de uma hipótese verdadeira é chamada "erro tipo I". A aceitação de uma hipótese falsa constitui um "erro tipo II".

As probabilidades desses dois tipos de erros são designados, respectivamente, por  $\alpha$  e  $\beta$ .

A probabilidade  $\alpha$  do erro do tipo I é denominada "nível de significância" do teste.

Os possíveis erros e acertos de um teste estão sintetizados abaixo:

		Realidade	
		$H_0$ verdadeira	$H_0$ falsa
Decisão	Aceitar $H_0$	Decisão correta ( $1 - \alpha$ )	Erro Tipo II ( $\beta$ )
	Rejeitar $H_0$	Erro Tipo I ( $\alpha$ )	Decisão Correta ( $1 - \beta$ )

Observe que o erro tipo I só poderá acontecer se for rejeitado  $H_0$  e o erro tipo II quando for aceito  $H_0$ .

## 9.2 Teste de significância

Os testes de Significância considera somente erros do tipo  $\alpha$ , pois são os mais usados em pesquisas educacionais, sócio-econômicas. . .

O procedimento para realização dos testes de significância é resumido nos seguintes passo:

- 1º) Enunciar as hipóteses  $H_0$  e  $H_1$ ;
- 2º) fixar o limite do erro  $\alpha$ , e identificar a variável do teste;
- 3º) com o auxílio das tabelas estatísticas, considerando  $\alpha$  e a variável do teste, determinar as RR (região de rejeição) e RA (região de aceitação) para  $H_0$ ;
- 4º) com os elementos amostrais, calcular o valor da variável do teste;
- 5º) concluir pela aceitação ou rejeição do  $H_0$  pela comparação do valor calculado no 4º passo com RA e RR.

### 9.2.1 Teste de significância para a média

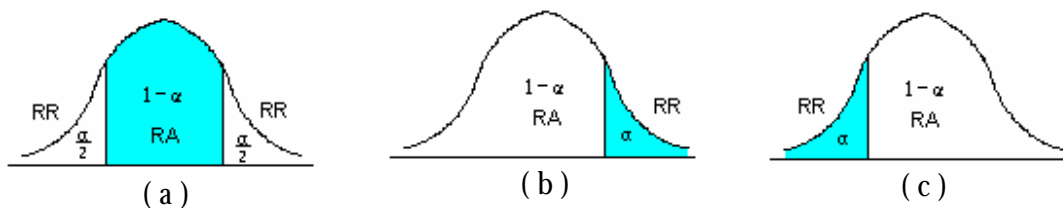
1. Enunciar as hipóteses:

$$H_0 : \mu = \mu_0 \qquad H_1 : \begin{cases} \mu \neq \mu_0 & (a) \\ \mu > \mu_0 & (b) \\ \mu < \mu_0 & (c) \end{cases}$$

2. Fixar  $\alpha$ . Admitindo:

- Se a variância populacional  $\sigma^2$  for conhecida, a variável teste será "Z" ( $n > 30$ );
- Se a variância populacional  $\sigma^2$  for desconhecida, a variável teste será "t" de Student com  $\varphi = n - 1$  ( $n \leq 30$ ).

3. Com o auxílio das tabelas "Z" e "t" determinar as regiões RA e RR;



4. Calcular o valor da variável:

$$Z_{\text{cal}} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \qquad t_{\text{cal}} = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

onde:  $\bar{X}$  = média amostral  
 $\mu_0$  = valor da hipótese nula

5. Conclusão para a situação (a)

- Se  $-Z_{\frac{\alpha}{2}} \leq Z_{\text{cal}} \leq +Z_{\frac{\alpha}{2}}$  ou  $-t_{\frac{\alpha}{2}} \leq t_{\text{cal}} \leq +t_{\frac{\alpha}{2}}$ , não se pode rejeitar  $H_0$ .

- Se  $Z_{\text{cal}} < -Z_{\frac{\alpha}{2}}$  ou  $Z_{\text{cal}} > +Z_{\frac{\alpha}{2}}$  ou  $t_{\text{cal}} < -t_{\frac{\alpha}{2}}$  ou  $t_{\text{cal}} > +t_{\frac{\alpha}{2}}$ , rejeita-se  $H_0$

OBS.: Para qualquer tipo de teste de significância devemos considerar:  
- Se a variável teste (calculada) cair dentro da região de aceitação (RA) não se pode rejeitar  $H_0$ ;  
- Se a variável teste (calculada) cair fora da região de aceitação (RA) rejeita-se  $H_0$

Ex.: Os dois registros dos últimos anos de um colégio, atestam para os calouros admitidos uma nota média de 115 pontos (teste vocacional). Para testar a hipótese de que a média de uma nova turma é a mesma, tirou-se, ao acaso, uma amostra de 20 notas, obtendo-se média 118 e desvio padrão 20. Admitir  $\alpha = 5\%$ , para efetuar o teste. (R = aceita-se  $H_0$ )

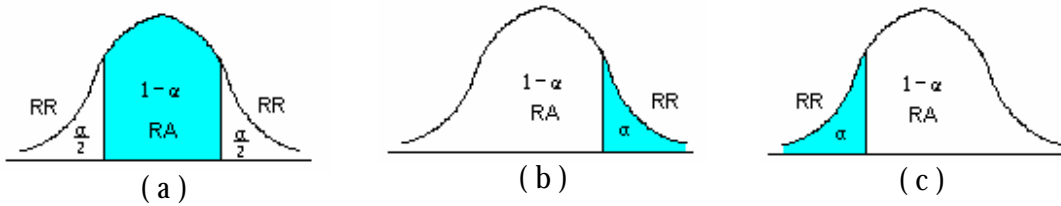
### 9.2.2 Teste de significância para a proporção

1. Enunciar as hipóteses:

$$H_0 : \pi = p_0 \qquad H_1 : \begin{cases} \pi \neq p_0 & \text{(a)} \\ \pi > p_0 & \text{(b)} \\ \pi < p_0 & \text{(c)} \end{cases}$$

2. Fixar  $\alpha$ . Escolhendo a variável normal padrão "Z";

3. Com o auxílio da tabela "Z" determinar as regiões RA e RR;



4. Calcular o valor da variável:

$$Z_{\text{cal}} = \frac{f - p_0}{\sqrt{\frac{p_0 \cdot q_0}{n}}} \quad \text{onde } f = \frac{X}{n}$$

onde:  $f$  = frequência relativa do evento na amostral  
 $X$  = característica dentro da amostra  
 $p_0$  = valor da hipótese nula

5. Conclusão para a situação (a)

- Se  $-Z_{\frac{\alpha}{2}} \leq Z_{\text{cal}} \leq +Z_{\frac{\alpha}{2}}$ , não se pode rejeitar  $H_0$ .

- Se  $Z_{\text{cal}} < -Z_{\frac{\alpha}{2}}$  ou  $Z_{\text{cal}} > +Z_{\frac{\alpha}{2}}$ , rejeita-se  $H_0$ .

Ex.: As condições de mortalidade de uma região são tais que a proporção de nascidos que sobrevivem até 60 anos é de 0,6 . Testar essa hipótese ao nível de 2%, se em 1000 nascimentos amostrados aleatoriamente, verificou-se 530 sobreviventes até 60 anos. (R = rejeita-se  $H_0$ )

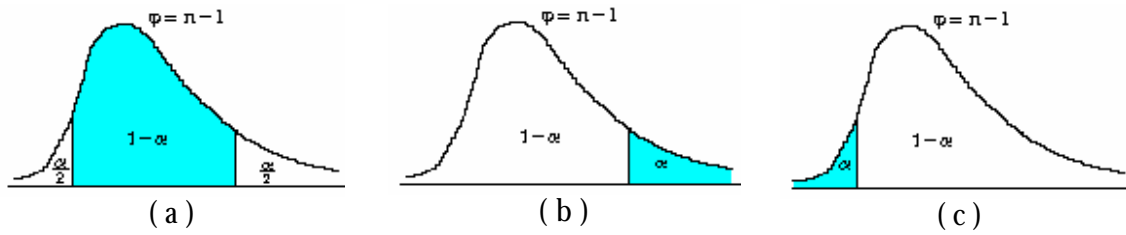
### 9.2.3 Teste de significância para a varância

1. Enunciar as hipóteses:

$$H_0 : \sigma^2 = \sigma_0^2 \quad H_1 : \begin{cases} \sigma^2 \neq \sigma_0^2 & \text{(a)} \\ \sigma^2 > \sigma_0^2 & \text{(b)} \\ \sigma^2 < \sigma_0^2 & \text{(c)} \end{cases}$$

2. Fixar  $\alpha$ . Escolhendo a variável qui-quadrado com  $\varphi = n - 1$ .

3. Com o auxílio da tabela " $\chi^2$ " determinar as regiões RA e RR;



Para (a) temos:  $\chi_{\text{inf}}^2 = \chi_{1-\frac{\alpha}{2}; \varphi}^2$ ,  $\chi_{\text{sup}}^2 = \chi_{\frac{\alpha}{2}; \varphi}^2$

Para (b) temos:  $\chi_{\text{sup}}^2 = \chi_{\alpha; \varphi}^2$

Para (c) temos:  $\chi_{\text{inf}}^2 = \chi_{1-\alpha; \varphi}^2$

4. Calcular o valor da variável:

$$\chi_{\text{cal}}^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

onde:  $S^2$  = variância amostral

$\sigma^2$  = valor da hipótese nula

5. Conclusão para a situação (a)

- Se  $\chi_{\text{inf}}^2 \leq \chi_{\text{cal}}^2 \leq \chi_{\text{sup}}^2$ , não se pode rejeitar  $H_0$ .

- Se  $\chi_{\text{cal}}^2 < \chi_{\text{inf}}^2$  ou  $\chi_{\text{cal}}^2 > \chi_{\text{sup}}^2$ , rejeita-se  $H_0$ .

Ex.: Para testar a hipótese de que a variância de uma população é 25, tirou-se uma amostra de 25 elementos obtendo-se  $S^2 = 18,3$ . Admitindo-se  $\alpha = 5\%$ , efetuar o teste de significância unicaudal a esquerda. (R = aceita-se  $H_0$ )

### 10.2.4 Teste de significância para igualdade de duas médias

1º caso) Se as variâncias populacionais  $\sigma^2$  forem conhecidas e supostamente iguais, independentes e normais, a variável teste será "Z" ( $n_1 + n_2 > 30$ );

1. Enunciar as hipóteses:

$$H_0 : \mu_1 = \mu_2 \text{ ou } \mu_1 - \mu_2 = d \qquad H_1 : \begin{cases} \mu_1 \neq \mu_2 \\ \mu_1 - \mu_2 \neq d \end{cases}$$



onde  $d$  é a diferença admitida entre as duas médias.

2. Fixar  $\alpha$ . Escolhendo a variável normal padrão "Z";
3. Com o auxílio da tabela "Z" determinar as regiões RA e RR;
4. Calcular o valor da variável:

$$Z_{\text{cal}} = \frac{(\bar{X}_1 - \bar{X}_2) - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

5. Conclusão:

Optar pela aceitação ou rejeição de  $H_0$ .

Ex.: Um fabricante de pneus faz dois tipos. Para o tipo A,  $\sigma = 2500$  Km, e para o tipo B,  $\sigma = 3000$  Km. Um taxi testou 50 pneus do tipo A e 40 pneus do tipo B, obtendo 24000 Km e 26000 Km de duração média dos respectivos tipos. Adotando um risco de 4% e que existe uma diferença admitida de 200 Km entre as marcas, testar a hipótese de que a vida média dos dois tipos é a mesma. (R = rejeita-se  $H_0$ )

2º caso) Se as variâncias populacionais  $\sigma^2$  forem desconhecidas, independentes e normais, a variável teste será "t" ( $n_1 + n_2 \leq 30$ ) com  $\varphi = n_1 + n_2 - 2$ ;

### a) As estimativas diferentes

1. Enunciar as hipóteses:

$$H_0 : \mu_1 = \mu_2 \text{ ou } \mu_1 - \mu_2 = d \qquad H_1 : \begin{cases} \mu_1 \neq \mu_2 \\ \mu_1 - \mu_2 \neq d \end{cases}$$

onde  $d$  é a diferença admitida entre as duas médias.

2. Fixar  $\alpha$ . Escolhendo a variável "t" de Student;
3. Com o auxílio da tabela "t" determinar as regiões RA e RR;
4. Calcular o valor da variável:

$$t_{\text{cal}} = \frac{(\bar{X}_1 - \bar{X}_2) - d}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

5. Conclusões:

Optar pela aceitação ou rejeição de  $H_0$ .

Ex.: Dois tipos de pneus foram testados sob as mesmas condições meteorológicas. O tipo A fabricado pela fábrica A, registrou uma média de 80.000 km rodados com desvio padrão de 5.000 km em 6 carros amostrados. O tipo B fabricado pela fábrica B, registrou uma média de 88.000 km com desvio padrão de 6.500 km em 12 carros amostrados. Adotando  $\alpha = 5\%$  testar a hipótese da igualdade das médias.

### b) As estimativas iguais

1. Enunciar as hipóteses:

$$H_0 : \mu_1 = \mu_2 \text{ ou } \mu_1 - \mu_2 = d \qquad H_1 : \begin{cases} \mu_1 \neq \mu_2 \\ \mu_1 - \mu_2 \neq d \end{cases}$$

onde  $d$  é a diferença admitida entre as duas médias.

2. Fixar  $\alpha$ . Escolhendo a variável "t" de Student;

3. Com o auxílio da tabela "t" determinar as regiões RA e RR;

4. Calcular o valor da variável:

$$t_{\text{cal}} = \frac{(\bar{X}_1 - \bar{X}_2) - d}{S_c \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \text{onde } S_c = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

5. Conclusões:

Optar pela aceitação ou rejeição de  $H_0$ .

Ex.: Dois tipos de tintas foram testados sob as mesmas condições meteorológicas. O tipo A registrou uma média de 80 min para secagem com desvio padrão de 5 min em cinco partes amostradas. O tipo B, uma média de 83 min com desvio padrão de 4 min em 6 partes amostradas. Adotando  $\alpha = 5\%$  testar a hipótese da igualdade das médias. ( $R = \text{aceita-se } H_0$ )

### 9.2.5 Teste de significância para igualdade de duas proporções

1. Enunciar as hipóteses:

$$H_0: \pi_1 = \pi_2 \quad H_1: \pi_1 \neq \pi_2$$

2. Fixar  $\alpha$ . Escolhendo a variável normal padrão "Z";

3. Com o auxílio da tabela "Z" determinar as regiões RA e RR;

4. Calcular o valor da variável:

$$Z_{\text{cal}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_{1-2} \cdot \hat{q}_{1-2}}{n}}} \quad \text{onde } \hat{p}_1 = \frac{X_1}{n_1}, \hat{p}_2 = \frac{X_2}{n_2}, \hat{p}_{1-2} = \frac{X_1 + X_2}{n_1 + n_2}$$

5. Conclusões:

Optar pela aceitação ou rejeição de  $H_0$ .

Ex.: Deseja-se testar se são iguais as proporções de homens e mulheres que lêem revista e se lembram do determinado anúncio. São os seguintes os resultados da amostra aleatória independente de homens e mulheres: Admita  $\alpha = 10\%$ .

Homens	Mulheres
$X_1 = 70$	$X_2 = 50$
$n_1 = 200$	$n_2 = 200$

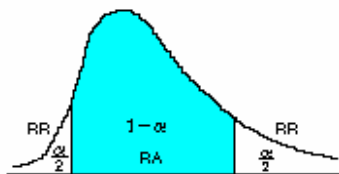
### 9.2.6 Teste de significância para igualdade de duas variâncias

1. Enunciar as hipóteses:

$$H_0: \sigma_1^2 = \sigma_2^2 \quad H_1: \sigma_1^2 \neq \sigma_2^2$$

2. Fixar  $\alpha$ . Escolhendo a variável "F" com  $\phi_1 = n_1 - 1$  graus de liberdade no numerador, e  $\phi_2 = n_2 - 1$  graus de liberdade no denominador.

3. Com o auxílio da tabela "F" determinar as regiões RA e RR;



$$F_{\text{sup}} = F_{\alpha}(\phi_1, \phi_2)$$

$$F_{\text{inf}} = F_{1-\alpha}(\phi_1, \phi_2) = \frac{1}{F_{\alpha}(\phi_2, \phi_1)}$$

4. Calcular o valor da variável:

$$F_{\text{cal}} = \frac{S_1^2}{S_2^2}$$

5. Conclusões:

Optar pela aceitação ou rejeição de  $H_0$ .

Ex.: Dois programas de treinamento de funcionários foram efetuados. Os 21 funcionários treinados no programa antigo apresentaram uma variância de 146 pontos em sua taxa de erro. No novo programa, 11 funcionários apresentaram uma variância de 200. Sendo  $\alpha = 10\%$ , pode-se concluir que a variância é diferente para os dois programas?

### 9.3 Exercícios

1) O crescimento da indústria da lagosta na Flórida nos últimos 20 anos tornou esse estado americano o 2º mais lucrativo centro industrial de pesca. Espera-se que uma recente medida tomada pelo governo das Bahamas, que proibiu os pescadores norte americanos de jogarem suas redes na plataforma continental desse país. Produza uma redução na quantidade de Kg de lagosta que chega aos Estados Unidos. De acordo com índices passados, cada rede traz em média 14 Kg de lagosta. Uma amostra de 20 barcos pesqueiros, recolhida após a vigência da nova lei, indicou os seguintes resultados, em Kg (Use  $\alpha = 5\%$ )

7.89	13.29	17.96	15.6	8.89
15.28	16.87	19.68	18.91	12.47
10.93	9.57	5.53	11.57	10.02
8.57	17.96	10.79	19.59	11.06

Estes dados mostram evidências suficientes de estar ocorrendo um decréscimo na quantidade média de lagostas pescadas por barco, que chega aos Estados Unidos, depois do decreto do governo das Bahamas? Teste considerando  $\alpha = 5\%$ . (bilateral)

2) Os resíduos industriais jogados nos rios, muitas vezes, absorvem oxigênio, reduzindo assim o conteúdo do oxigênio necessário à respiração dos peixes e outras formas de vida aquática. Uma lei estadual exige um mínimo de 5 p.p.m. (Partes por milhão) de oxigênio dissolvido, a fim de que o conteúdo de oxigênio seja suficiente para manter a vida aquática. Seis amostras de água retiradas de um rio, durante a maré baixa, revelaram os índices (em partes por milhão) de oxigênio dissolvido:

4.9      5.1      4.9      5.5      5.0      4.7

Estes dados são evidência para afirmar que o conteúdo de oxigênio é menor que 5 partes por milhão? Teste considerando  $\alpha = 5\%$  e  $\alpha = 1\%$ .

- 3) A Debug Company vende um repelente de insetos que alega ser eficiente pelo prazo de 400 horas no mínimo. Uma análise de 90 itens aleatoriamente inspecionados acusou uma média de eficiência de 380 horas.
- a) Teste a afirmativa da companhia, contra a alternativa que a duração é inferior a 400 horas, ao nível de 1%, seu desvio padrão é de 60 horas.  
b) Repita o teste, considerando um desvio padrão populacional de 90 horas.
- 4) Ao final de 90 dias de um dieta alimentar envolvendo 32 pessoas, constatou-se o seguinte ganhos médio de peso 40 g, e desvio padrão de 1,378g.
- a) Supondo que o ganho de peso médio dessas pessoas é de 45 g, teste a hipótese para  $\alpha = 5\%$ , se esse valor é o mesmo (bilateral)  
b) Supondo que a variância dessas pessoas é de 1.8 g<sup>2</sup>, teste a hipótese para  $\alpha = 5\%$ , se esse valor é o mesmo (bilateral).
- 5) Uma pesquisa feita alega que 15% dos pessoas de uma determinada região sofrem de cegueira aos 70 anos. Numa amostra aleatória de 60 pessoas acima de 70 anos constatou-se que 12 pessoas eram cegas. Teste a alegação para  $\alpha = 5\%$  contra  $p > 15\%$ .
- 6) A tabela abaixo mostra a quantidade de pessoas que obtiveram o melhor efeito perante a aplicação de duas drogas:

Sexo	Droga A		Droga B	Total
	muscular	intravenosa	muscular	
Masculino	21	10	22	53
Feminino	20	12	25	57
Total	41	22	47	110

- a) Testar a hipótese de que a proporção de homens submetidos a droga A é de 35%, sendo  $\alpha = 3\%$ .  
b) Testar a hipótese de que a proporção dos adultos que tiveram melhor resultado nas aplicações musculares é de 85%, usando  $\alpha = 2\%$ .  
c) Testar a hipótese de que a proporção de mulheres é de 50%, usando  $\alpha = 5\%$ .  
d) Testar a hipótese de que a proporção de pessoas submetidos a droga A é de 65%, usando  $\alpha = 4\%$ .
- 7) Um processo de fabricação de arame de aço dá um produto com resistência média de 200 psi. O desvio padrão é de 20 psi. O engenheiro de controle de qualidade deseja elaborar um teste que indique se houve ou não variação na média do

processo, usando uma amostra de 25 arames obteve-se uma média de 285 psi. Use um nível de significância de 5%. Suponha normal a população das resistências.

- 8) A DeBug Company vende um repelente para insetos que alega ser eficiente pelo prazo de 400 horas no mínimo. Uma análise de 9 itens escolhidos aleatoriamente acusou uma média de eficiência de 380 horas. a) Teste a alegação da companhia, contra a alternativa que a duração é inferior a 400 horas no mínimo ao nível de 1%, e o desvio padrão amostral é de 60 horas. b) Repita o item (a) sabendo que o desvio padrão populacional é de 90 horas.
- 9) Um laboratório de Análises Clínicas realizou um teste de impurezas em 9 porções de um determinado composto. Os valores obtidos foram: 10,32; 10,44; 10,56; 10,60; 10,63; 10,67; 10,7; 10,73; 10,75 mg. a) estimar a média e a variância de impurezas entre as porções. b) Testar a hipótese de que a média de impureza é de 10,4, usando  $\alpha = 5\%$ . c) Testar a hipótese de que a variância é um usando  $\alpha = 5\%$ .
- 10) Uma experiência tem mostrado que 40% dos estudantes de uma Universidade reprovam em pelo menos 5 disciplinas cursada na faculdade. Se 40 de 90 estudantes fossem reprovados em mais de 5 disciplinas, poderíamos concluir quanto a proporção populacional, usando  $\alpha = 1\%$ .
- 11) Para verificar a eficácia de uma nova droga foram injetados doses em 72 ratos, obtendo-se a seguinte tabela:

Sexo	Tamanho da Amostra	Variância
Machos	41	43.2
Fêmeas	31	29.5

Testar a igualdade das duas variâncias usando  $\alpha = 10\%$ .

12) Sendo

Amostra 1	$n_1 = 60$	$\bar{X}_1 = 5.71$	$\sigma_1^2 = 43$
Amostra 2	$n_2 = 35$	$\bar{X}_2 = 4.12$	$\sigma_2^2 = 28$

- a) Testar a igualdade das duas médias usando  $\alpha = 4\%$ .  
 b) Testar a igualdade das duas variâncias usando  $\alpha = 5\%$

- 13) Na tabela abaixo estão registrados os índices de vendas em 6 supermercados para os produtos concorrentes da marca A e marca B. Testar a hipótese de que a diferença das médias no índice de vendas entre as marcas é zero, usando  $\alpha = 5\%$ .

Supermercado	Marca A	Marca B
1	14	4
2	20	16
3	2	28
4	11	9
5	5	31
6	12	10

- 14) Da população feminina extraiu-se uma amostra resultando:

Renda (em \$1000)	10  --- 25	25  --- 40	40  --- 55	55  --- 70	70  ---85
Nº de mulheres	7	12	10	6	4

da população masculina retirou-se uma amostra resultando:

Renda (em \$1000)	10  --- 25	25  --- 40	40  --- 55	55  --- 70	70  ---85
Nº de homens	8	15	12	7	3

- a) Testar ao nível de 10% a hipótese de que a diferença entre a renda média dos homens e das mulheres valha \$ 5000.
- b) Testar ao nível de 5% a hipótese de que a diferença entre as variâncias valha 0.
- 15) Uma empresa de pesquisa de opinião seleciona, aleatoriamente, 300 eleitores de São Paulo e 400 do Rio de Janeiro, e pergunta a cada um se votará ou não num determinado candidato nas próximas eleições. 75 eleitores de SP e 120 do RJ responderam afirmativo. Há diferença entre as proporções de eleitores favoráveis ao candidato naqueles dois estados? use  $\alpha = 5\%$ .
- 16) Estão em teste dois processos para fechar latas de comestíveis. Numa seqüência de 1000 latas, o processo 1 gera 50 rejeições, enquanto o processo 2 acusa 200 rejeições. Pode ao nível de 5%, concluir que os dois processos sejam diferentes?

#### 9.4 Teste do Qui-quadrado $\chi^2$

Anteriormente foi testado hipóteses referentes a um parâmetro populacional ou mesmo a comparação de dois parâmetros, ou seja, são os testes paramétricos. Agora será testado aqueles que não dependem de um parâmetro populacional, nem de suas respectivas estimativas. Este teste é chamado de teste do Qui-quadrado.

O teste do Qui-quadrado é usado quando se quer comparar frequências observadas com frequências esperadas. Divide-se em três tipos:

- Teste de adequação do ajustamento
- Teste de Aderência
- Teste de Independência (Tabela de Contingência)

### 9.4.1 Procedimentos para a realização de um teste Qui-quarado ( $\chi^2$ )

1ª) Determinação das Hipóteses:

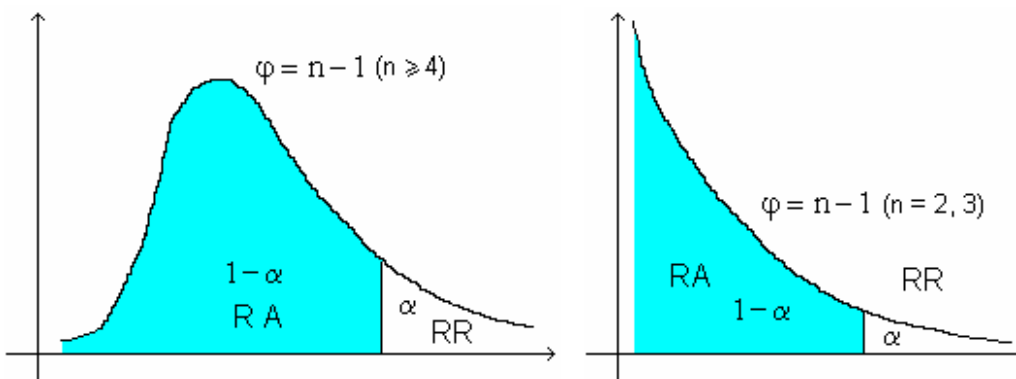
$$H_0: F_o = F_e \quad H_1: F_o \neq F_e$$

2ª) Escolha do Nível de Significância ( $\alpha$ )

3ª) Estatística Calculada 
$$\chi_{cal}^2 = \sum_{i=1}^k \frac{(F_o - F_e)^2}{F_e}$$

4ª) Estatística Tabelada:  $\chi_{tab}^2 = \chi_{\varphi, \alpha}^2$

5ª) Comparar o  $\chi_{cal}^2$  com  $\chi_{tab}^2$  e concluir:



6ª) Conclusão:

Se  $\chi_{cal}^2 \leq \chi_{tab}^2 \rightarrow$  aceita-se  $H_0$

Se  $\chi_{cal}^2 > \chi_{tab}^2 \rightarrow$  rejeita-se  $H_0$



### 9.4.2 Teste de adequação do ajustamento

Suponhamos uma amostra de tamanho  $n$ . Sejam  $E_1, E_2, \dots, E_k$ , um conjunto de eventos possíveis da amostra.

Eventos	Freq. Obs.
$E_1$	$F_{o1}$
$E_2$	$F_{o2}$
$\vdots$	$\vdots$
$E_k$	$F_{ok}$
Total	$n$

Este teste é indicado para verificar se as frequências observadas dos  $k$  eventos ( $k$  classes em que a variável é dividida) concordam ou não com as frequências teóricas esperadas.

As frequências esperadas ( $F_{ei}$ ) são obtidas multiplicando-se o número total de elementos pela proporção teórica da classe  $i$  ( $n \cdot p_i$ ).

Para encontrar o  $\chi^2_{cal}$ , necessita-se do nível de Significância ( $\alpha$ ) e dos graus de liberdade ( $\phi$ ), os quais podem ser obtidos da seguinte forma:

1º)  $\phi = k - 1$ , quando as frequências esperadas puderem ser calculadas sem que façam estimativas dos parâmetros populacionais a partir das distribuições amostrais.

2º)  $\phi = k - 1 - m$ , quando para a determinação das frequências esperadas,  $m$  parâmetros tiverem suas estimativas calculadas a partir das distribuições amostrais. Pearson mostrou que, se o modelo testado for verdadeiro e se todas as  $F_{ei} \leq 5\%$ , estas deverão ser fundidas às classes adjacentes.

Ex1.: Deseja-se testar se o número de acidentes numa rodovia se distribui igualmente pelos dias da semana. Para tanto foram levantados os seguintes dados ( $\alpha = 5\%$ ).

Dia da Semana	Dom	Seg	Ter	Qua	Qui	Sex	Sab
Nº de Acidentes	33	26	21	22	17	20	36

Resolução:

$H_0$ : As frequências são iguais em todos os dias

$H_1$ : As frequências são diferentes em todos os dias

Dia da Semana	Dom	Seg	Ter	Qua	Qui	Sex	Sab	k = 7
Nº de Acidentes	33	26	21	22	17	20	36	175
$p_i$	1/7	1/7	1/7	1/7	1/7	1/7	1/7	---
$Fe = n \cdot p_i$	25	25	25	25	25	25	25	---

$$\chi^2_{\text{calc}} = \frac{(33-25)^2}{25} + \frac{(26-25)^2}{25} + \dots + \frac{(36-25)^2}{25} = 12$$

$$\chi^2_{\text{cal}} = 12 \quad \chi^2_{\text{tab}} = \chi^2_{5\%,6} = 12.5$$

Conclusão: Como o  $\chi^2_{\text{cal}} < \chi^2_{\text{tab}}$ , aceita-se  $H_0$ , ou seja, para  $\alpha = 5\%$ , as frequências podem ser iguais.

Ex2.: O número de livros emprestados por uma biblioteca durante certa semana está a seguir. Teste a hipótese que o número de livros emprestados não dependem do dia da semana, com  $\alpha = 1\%$ .

Dia da Semana	Seg	Ter	Qua	Qui	Sex
Nº de Livros	110	135	120	146	114

$$R = \chi^2_{\text{cal}} = 7.29 ; \chi^2_{\text{tab}} = \chi^2_{1\%,4} = 13.28$$

### 9.4.3 Teste de aderência

É usada a estatística  $\chi^2$  quando deseja-se testar a natureza da distribuição amostral. Por exemplo, quando se quer verificar se a distribuição amostral se ajusta a um determinado modelo de distribuição de probabilidade (Normal, Poisson, Binomial, ...), ou seja, verifica-se a boa ou má, aderência dos dados da amostra do modelo.

Ex1.: O número de defeitos por unidade de uma amostra de 100 aparelhos de TV produzidos por uma linha de montagem apresentou a seguinte distribuição:

Nº de Defeitos	0	1	2	3	4	5	6	7
Nº de Aparelhos	25	35	18	13	4	2	2	1

Verificar se o número de defeitos por unidade segue razoavelmente uma distribuição de Poisson, com  $\alpha = 5\%$ .

Resolução:

$H_0$ : A distribuição do nº de defeitos/unidade segue uma Poisson

$H_1$ : A distribuição do nº de defeitos/unidade não segue uma Poisson

Nº de Defeitos	0	1	2	3	4	5	6	7	n
Nº de Aparelhos	25	35	18	13	4	2	2	1	100
$p_i$	0.212	0.329	0.255	0.132	0.051	0.016	0.004	0.001	1.000
$Fe = n \cdot p_i$	21.2	32.9	25.5	13.2	5.1	1.6*	0.4*	0.1*	100

Para calcular  $p_i$ , temos:  $p_i = \frac{e^{-\mu} \mu^i}{i!}$  (Distribuição de Poisson)

$$\mu = \bar{X} = \frac{\sum_{i=1}^n X_i \cdot f_i}{n} = 1.55$$

$$p_0 = \frac{e^{-1.55} (1.55)^0}{0!} = 0.212 \quad p_4 = \frac{e^{-1.55} (1.55)^4}{4!} = 0.051$$

$$p_1 = \frac{e^{-1.55} (1.55)^1}{1!} = 0.329 \quad p_5 = \frac{e^{-1.55} (1.55)^5}{5!} = 0.016$$

$$p_2 = \frac{e^{-1.55} (1.55)^2}{2!} = 0.255 \quad p_6 = \frac{e^{-1.55} (1.55)^6}{6!} = 0.004$$

$$p_3 = \frac{e^{-1.55} (1.55)^3}{3!} = 0.132 \quad p_7 = \frac{e^{-1.55} (1.55)^6}{6!} = 0.001$$

\* =  $F_{ei} \leq 5\%$ , logo deve ser agrupada com a classe adjacente.

$$\chi_{\text{calc}}^2 = \frac{(25 - 21.2)^2}{21.2} + \frac{(35 - 32.9)^2}{32.9} + \dots + \frac{(9 - 7.2)^2}{\underbrace{7.2}_{4-5-6-7}} = 3.474$$

$$\varphi = k - 1 - m = 5 - 1 - 1 = 3 \quad \begin{cases} k = 5 \text{ (número de classes após agrupamento)} \\ m = 1 \text{ (número de estimadores usados) } (\bar{X}) \end{cases}$$

$$\chi_{\text{tab}}^2 = \chi_{5\%,3}^2 = 7.81$$

Conclusão: Como o  $\chi_{\text{cal}}^2 \leq \chi_{\text{tab}}^2$ , aceita-se  $H_0$ , ou seja, para  $\alpha = 5\%$ , a distribuição do número de defeitos/unidade pode seguir uma Distribuição de Poisson.

Ex2.: Verificar se os dados das distribuição das alturas de 100 estudantes do sexo feminino se aproxima de uma distribuição normal, com  $\alpha = 5\%$ .

Altura (cm)	Nº de Estudantes	Transf. "Z"	Prob (área)	Fe = n . p <sub>i</sub>
150  --- 156	4	∞  --- -1.94	0.026 *	2.6
156  --- 162	12	-1.94  --- -1.04	0.123	12.3
162  --- 168	22	-1.04  --- -0.14	0.295	29.5
168  --- 174	40	-0.14  --- 0.76	0.332	33.2
174  --- 180	20	0.75  --- 1.66	0.175	17.5
180  --- 186	2	1.66  --- +∞	0.048 *	4.8
k = 6	100		1.000	100.0

Resolução:

$H_0$ : A distribuição da altura das estudantes do sexo feminino é normal

$H_1$ : A distribuição da altura das estudantes do sexo feminino não é normal

Para calcular  $p_i$ , temos:  $Z_i = \frac{x - \mu}{S}$  (Distribuição Normal Padrão)

$$\mu = \bar{X} = \frac{\sum_{i=1}^n X_i \cdot f_i}{n} = 168.96$$

$$S = \frac{\sum_{i=1}^n f_i \cdot (X_i - \bar{X})^2}{n - 1} = 6.67$$

$$Z_1 = \frac{156 - 168.96}{6.67} = -1.94$$

$$Z_4 = \frac{174 - 168.96}{6.67} = 0.76$$

$$Z_2 = \frac{162 - 168.96}{6.67} = -1.04$$

$$Z_5 = \frac{180 - 168.96}{6.67} = 1.65$$

$$Z_3 = \frac{168 - 168.96}{6.67} = -0.14$$

$$\chi^2_{\text{calc}} = \frac{((4+12) - (2.6+12.3))^2}{14.9} + \frac{(22 - 29.5)^2}{29.5} + \frac{(40 - 33.2)^2}{33.2} + \frac{((20+2) - (17.5+4.8))^2}{22.3} = 3.38$$

$$\varphi = k - 1 - m = 4 - 1 - 2 = 1 \begin{cases} k = 4 \text{ (número de classes após agrupamento)} \\ m = 2 \text{ (número de estimadores usados) } (\bar{X}) (S) \end{cases}$$

$$\chi^2_{\text{tab}} = \chi^2_{5\%,1} = 3.84$$

Conclusão: Como o  $\chi^2_{\text{cal}} \leq \chi^2_{\text{tab}}$ , aceita-se  $H_0$ , ou seja, para  $\alpha = 5\%$ , a distribuição da altura das estudantes do sexo feminino é normal.

### 9.4.4 Tabelas de contingência – Teste de independência

Uma importante aplicação do teste  $\chi^2$  ocorre quando se quer estudar a relação entre duas ou mais variáveis de classificação. A representação das frequências observadas, nesse caso, pode ser feita por meio de uma tabela de contingência.

$H_0$ : As variáveis são independentes (não estão associadas)

$H_1$ : As variáveis não são independentes (estão associadas)

O número de graus de liberdade é dado por:  $\varphi = (L - 1) (C - 1)$ , onde L é o número de linhas e C o número de colunas da tabela de contingência.

$$\chi^2_{\text{cal}} = \sum_{i=1}^k \frac{(F_o - F_e)^2}{F_e} \quad F_{eij} = \frac{(\text{soma da linha } i)(\text{soma da coluna } j)}{\text{total de observações}}$$

Obs.: Em tabelas 2 x 2 temos 1 grau de liberdade, por isso utiliza-se a correção de Yates, onde para  $n \geq 50$  pode ser omitida.

O teste do  $\chi^2$  não é indicado em tabelas 2 x 2 nos seguintes casos:

- i) quando alguma frequência esperada for menor que 1;
- ii) quando a frequência total for menor que 20 ( $n \leq 20$ );
- iii) quando a frequência total ( $20 \leq n \leq 40$ ) e algumas frequências esperadas for menor que 5. Neste caso aplica-se o teste exato de Fischer.

Ex.: Verifique se há associação entre os níveis de renda e os municípios onde foram pesquisados 400 moradores. Use  $\alpha = 1\%$ .

Município	Níveis de Renda			
	A	B	C	D
X	28	42	30	24
Y	44	78	78	76

$H_0$ : As variáveis são independentes

$H_1$ : As variáveis são dependentes

	A	B	C	D	Total ( $\ell$ )
X	28	42	30	24	124
Y	44	78	78	76	276
Total (c)	72	120	108	100	400

$$F_{e11} = \frac{124 \times 72}{400} = 22.32$$

$$F_{e21} = \frac{276 \times 72}{400} = 49.68$$

$$F_{e12} = \frac{124 \times 120}{400} = 37.2$$

$$F_{e22} = \frac{276 \times 120}{400} = 82.8$$

$$F_{e13} = \frac{124 \times 108}{400} = 33.48$$

$$F_{e23} = \frac{276 \times 108}{400} = 74.52$$

$$F_{e14} = \frac{124 \times 100}{400} = 31.0$$

$$F_{e24} = \frac{276 \times 100}{400} = 69$$

$$\chi^2_{\text{calc}} = \frac{(28 - 22,32)^2}{22,32} + \frac{(42 - 37,2)^2}{37,2} + \dots + \frac{(76 - 69)^2}{69} = 5,81$$

$$\chi^2_{\text{tab}} = \chi^2_{1\%;(2-1)(4-1)} = \chi^2_{1\%;3} = 11,34$$

Conclusão: Como o  $\chi^2_{\text{calc}} \leq \chi^2_{\text{tab}}$ , aceita-se  $H_0$ , ou seja, para  $\alpha = 1\%$ , as variáveis são independentes.

Uma medida do grau de relacionamento, associação ou dependência das classificações em uma tabela de contingência é dada pelo coeficiente de contingência.

$$C = \sqrt{\frac{\chi^2_{\text{calc}}}{\chi^2_{\text{calc}} + n}}$$

Quanto maior o valor de C, maior o grau de associação, O máximo valor de C dependerá do número de linhas e colunas da tabela e pode variar de 0 (independência) a 1 (dependência total).

## 9.5 Exercícios

- 1) Verificar se os dados se ajustam a uma distribuição de Poisson,  $\alpha = 2.5\%$ .

Nº de Acidentes	0	1	2	3	4	5
Nº de Dias	25	19	10	9	4	3

- 2) Testar para  $\alpha = 5\%$  se há alguma relação entre as notas escolares e o salário.

Salários	Notas Escolares		
	Alta	Baixa	Média
Alto	18	5	17
Médio	26	16	38
Baixo	6	9	15

R = aceita  $H_0$

3) Determine o valor do coeficiente de contingência considerando os dados:  $\alpha = 1\%$ .

Sexo	Partido	
	A	B
Masculino	50	72
Feminino	29	35

$C = 4\%$

4) Com o objetivo de investigar a relação entre a situação do emprego no momento em que se aprovou um empréstimo e saber se o empréstimo está, agora, pago ou não, o gerente de uma financeira selecionou ao acaso 100 clientes obtendo os resultados da tabela. Teste a hipótese nula de que a situação de emprego e a de empréstimo são variáveis independentes, com  $\alpha = 5\%$ .

Estado Atual do Empréstimo	Situação de Emprego	
	Empregado	Desempregado
Em mora	10	8
Em dia	60	22

$R = \text{Aceita } H_0$

5) A tabela indica o número médio de acidentes por mil homens/hora da amostra de 50 firmas obtidas de uma indústria específica. A média desta distribuição é de 2.32 e o desvio padrão de 0.42. Teste a hipótese nula de que as frequências observadas seguem uma distribuição normal, com  $\alpha = 5\%$ .

Nº de Acidentes	Nº de Firmas
1.5 --- 1.7	3
1.8 --- 2.0	12
2.1 --- 2.3	14
2.4 --- 2.6	9
2.7 --- 2.9	7
3.0 --- 3.2	5
Total	50

$R = \text{Aceita } H_0$

6) Verifique se a distribuição se ajusta a normal. ( $\alpha = 5\%$ ), onde 200 casas de aluguel foram pesquisadas, obtendo-se a seguinte distribuição:

Classes (\$)	freq. observada
250   --- 750	2
750   --- 1250	10
1250   --- 1750	26
1750   --- 2250	45
2250   --- 2750	60
2750   --- 3250	37
3250   --- 3750	13
3750   --- 4250	5
4250   --- 4750	2
Total	200

# 10 REGRESSÃO E CORRELAÇÃO

## 10.1 Introdução

Muitas vezes é de interesse estudar-se um elemento em relação a dois ou mais atributos ou variáveis simultaneamente.

Nesses casos presume-se que pelo menos duas observações são feitas sobre cada elemento da amostra. A amostra consistirá, então, de pares de valores, um valor para cada uma das variáveis, designadas, X e Y. Um indivíduo “i” qualquer apresenta o par de valores  $(X_i ; Y_i)$ .

Objetivo visado quando se registra pares de valores (observações) em uma amostra, é o estudo das relações entre as variáveis X e Y.

Para a análise de regressão interessam principalmente os casos em que a variação de um atributo é sensivelmente dependente do outro atributo.

O problema consiste em estabelecer a função matemática que melhor exprime a relação existente entre as duas variáveis. Simbolicamente a relação é expressa por uma equação de regressão e graficamente por uma curva de regressão.

## 10.2 Definição

Constitui uma tentativa de estabelecer uma equação matemática linear que descreva o relacionamento entre duas variáveis (uma dependente e outra independente).

A equação de regressão tem por finalidade **ESTIMAR** valores de uma variável, com base em valores conhecidos da outra.

Ex.: Peso x Idade; Vendas x Lucro; Nota x Horas de Estudo

## 10.3 Modelo de Regressão

$$\hat{y}_i = \alpha x_i + \beta + \varepsilon_i$$

$\hat{y}_i \Rightarrow$  valor estimado (variável dependente);

$x_i \Rightarrow$  variável independente;

$\alpha \Rightarrow$  coeficiente de regressão (coeficiente angular);

$\beta \Rightarrow$  coeficiente linear;

$\varepsilon_i \Rightarrow$  resíduo.



### 10.3.1 Pressuposições

Na regressão, os valores de "y" são estimados com base em valores dados ou conhecidos de "x", logo a variável "y" é chamada variável dependente, e "x" variável independente.

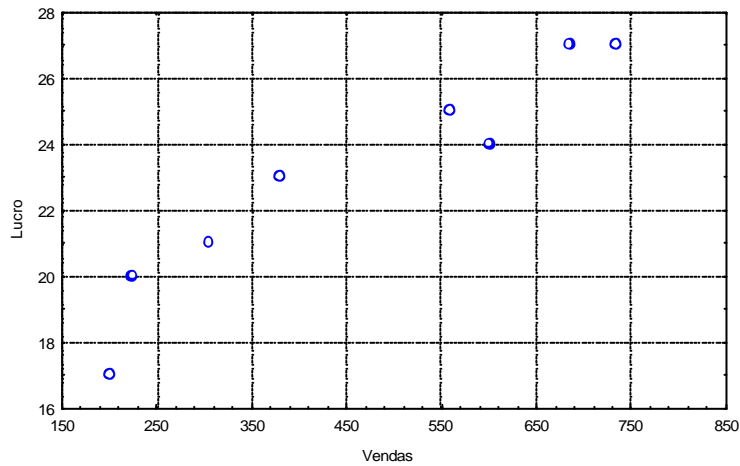
- A relação entre X e Y é linear (os acréscimos em X produzem acréscimos proporcionais em Y e a razão de crescimento é constante).
- Os valores de X são fixados arbitrariamente (X não é uma variável aleatória).
- Y é uma variável aleatória que depende entre outras coisas dos valores de X.
- $\varepsilon_i$  é o erro aleatório, portanto uma variável aleatória com distribuição normal, com média zero e variância  $\sigma^2$ . [ $\varepsilon_i \sim N(0, \sigma^2)$ ].  $\varepsilon_i$  representa a variação de Y que não é explicada pela variável independente X.
- Os erros são considerados independentes.

Ex.: Vendas (x 1000) X Lucro (x100)

Obs.	1	2	3	4	5	6	7	8
Vendas	201	225	305	380	560	600	685	735
Lucro	17	20	21	23	25	24	27	27

### 10.3.1 Gráfico (Diagrama de Dispersão)

Tem como finalidade ajudar na decisão se uma reta descreve adequadamente ou não o conjunto de dados.



Pelo gráfico podemos observar que a possível reta de regressão terá um coeficiente de regressão (coeficiente linear) positivo.

## 10.4 Método para estimação dos parâmetros a e b

As estimativas dos parâmetros  $\alpha$  e  $\beta$  dadas por "a" e "b", serão obtidas a partir de uma amostra de n pares de valores  $(x_i, y_i)$  que correspondem a n pontos no diagrama de dispersão. Exemplo:

O método mais usado para ajustar uma linha reta para um conjunto de pontos  $(x_1, y_1), \dots, (x_n, y_n)$  é o **Método de Mínimos Quadrados**:

O método dos mínimos quadrados consiste em adotar como estimativa dos parâmetros os valores que minimizem a soma dos quadrados dos desvios.

### 10.4.1 Características

1a) A soma dos desvios verticais dos pontos em relação a reta é zero;

2a) A soma dos quadrados desses desvios é mínima.

Os valores de "a" e "b" da reta de regressão  $\hat{y} = a x + b$  serão:

$$a = \frac{n \sum_{i=1}^n xy - \sum_{i=1}^n x \sum_{i=1}^n y}{n \sum_{i=1}^n x^2 - \left( \sum_{i=1}^n x \right)^2} = \frac{S_{xy}}{S_{xx}}$$

$$b = \bar{y} - a \bar{x}$$

Para cada par de valores  $(x_i, y_i)$  podemos estabelecer o desvio:

$$e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$$

Para facilitar os cálculos da reta de regressão, acrescentamos três novas colunas na tabela dada.

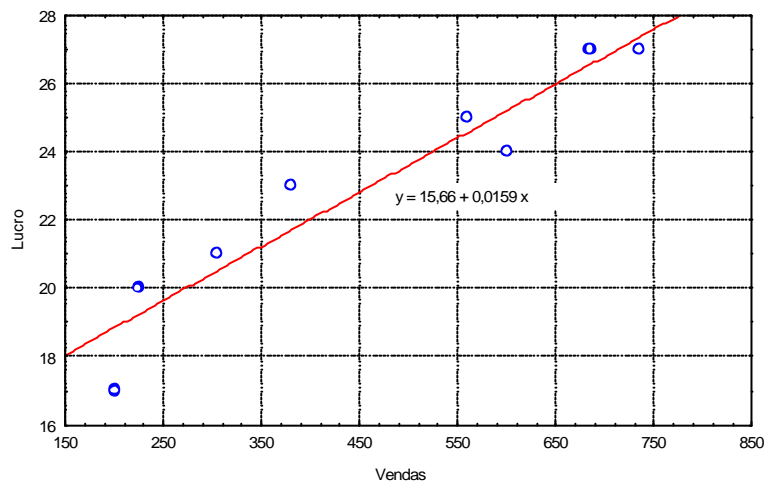
Obs.	Vendas $X_i$	Lucro $Y_i$	$X_i^2$	$Y_i^2$	$X_i \cdot Y_i$
1	201	17	40401	289	3417
2	225	20	50625	400	4500
3	305	21	93025	441	6405
4	380	23	144400	529	8740
5	560	25	313600	625	14000
6	600	24	360000	576	14400
7	685	27	469225	729	18495
8	735	27	540225	729	19845
$\Sigma$	3691	184	2011501	4318	89802

$$a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} = \frac{8(89802) - (3691)(184)}{8(2011501) - (3691)^2} = 0,0159$$

$$b = \bar{y} - a \bar{x} = 23 - (0,159)(461,38) = 15,66$$

$$\hat{y} = 0,0159 x + 15,66$$

### Saída do Statistica



Partindo da reta de regressão podemos afirmar que para uma venda de 400 mil podemos obter um lucro de  $\hat{y} = (0.0159) (400.000) + 15.66 = 22$  mil .

### Saídas do Statistica

Predicting Values for (regress.sta)			
Continue...	variable: VENDAS		
variable	B-Weight	Value	B-Weight * Value
LUCRO	57,08140	400,0000	22832,56
Intercept			-851,50
Predictd			21981,06
-95,0%CL			15142,68
+95,0%CL			28819,44

Obs.: Para qualquer tipo de equação de regressão devemos ter muito cuidado para não extrapolar valores para fora do âmbito dos dados. O perigo da extrapolação para fora dos dados amostrais, é que a mesma relação possa não mais ser verificada.

### 10.5 Decomposição da variância Total

A dispersão da variação aleatória “y” pode ser medida através da soma dos quadrados dos desvios em relação a sua média  $\bar{y}$ . Essa soma de quadrados será denominada Soma de Quadrados Total (SQTotal).

$$SQTotal = \sum_{i=1}^n (y_i - \bar{y})^2$$

A SQTotal pode ser decomposta da seguinte forma:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Essa relação mostra que a variação dos valores de Y em torno de sua média pode ser dividida em duas partes: uma  $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  que é explicada pela regressão e outra  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ , devido ao fato de que nem todos os pontos estão sobre a reta de regressão, que é a parte “não explicada” pela regressão ou variação residual.

Assim:

$$SQ. Total = SQRegressão + SQResíduo$$

A estatística definida por  $r^2 = \frac{SQRegressão}{SQTotal}$ , e denominada coeficiente de determinação, indica a proporção ou percentagem da variação de Y que é “explicada” pela regressão.

Note que  $0 \leq r^2 \leq 1$ .

Fórmulas para cálculo:

$$SQTotal = \sum_{i=1}^n (y_i - \bar{y})^2 = n \sum_{i=1}^n y_i^2 - \left( \sum_{i=1}^n y_i \right)^2,$$

com (n - 1) graus de liberdade.

$$SQ \text{ Regressão} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b \left( n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \right),$$

com (n - 2) graus de liberdade.

## 10.6 Análise de Variância da Regressão

A Soma de Quadrados da Regressão (SQRegressão), segue uma distribuição de  $\chi^2$  (qui-quadrado) com (1) grau de liberdade, enquanto que a Soma de Quadrados do Resíduo (SQResíduos) segue a mesma distribuição, porém com (n - 2) graus de liberdade. Portanto, o quociente

$$\frac{SQRegress\tilde{a}o/1}{SQResiduo/n - 2} = \frac{QMRegress\tilde{a}o}{QMResiduo},$$

segue uma distribuição F de Snedecor com 1 e (n - 2) graus de liberdade.

Esse fato nos permite empregar a distribuição F de Snedecor para testar a significância da regressão, através da chamada Análise de Variância, sintetizada no quadro abaixo:

Análise de Variância

Causas de Variação	G.L.	SQ	QM	F
Regressão	1	SQRegressão	$\frac{QMRegress\tilde{a}o}{1}$	$\frac{QMRegress\tilde{a}o}{QMResiduo}$
Resíduo	n - 2	SQResíduo	$\frac{QMResiduo}{n - 2}$	---
Total	n - 1	SQTotal	---	---

onde QM representa Quadrado Médio e é obtido pela divisão da Soma de Quadrados pelos respectivos graus de liberdade.

Para testar a significância da regressão, formula-se as seguintes hipóteses:

$H_0: \beta=0$  contra  $H_1: \beta\neq 0$ , onde  $\beta$  representa o coeficiente de regressão paramétrico.

Se o valor de F, calculado a partir do quadro anterior, superar o valor teórico de F com 1 e (n - 2) graus de liberdade, para o nível de significância  $\alpha$ , rejeita-se  $H_0$  e conclui-se que a regressão é significativa.

Se  $F_{calc.} > F_{\alpha [1, (n-2)]}$ , rejeita-se  $H_0$ .

Para o exemplo anterior:

Obs.	Vendas $X_i$	Lucro $Y_i$	$X_i^2$	$Y_i^2$	$X_i \cdot Y_i$
1	201	17	40401	289	3417
2	225	20	50625	400	4500
3	305	21	93025	441	6405
4	380	23	144400	529	8740
5	560	25	313600	625	14000
6	600	24	360000	576	14400
7	685	27	469225	729	18495
8	735	27	540225	729	19845
$\Sigma$	3691	184	2011501	4318	89802

$$\hat{y}_i = 0,0159x_i + 15,66$$

$$SQ_{\text{Regressão}} = a \left[ n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \right]$$

$$SQ_{\text{Regressão}} = 0,0159[8(89802) - (3691)(184)] = 624,42$$

$$SQ_{\text{Total}} = n \sum_{i=1}^n y_i^2 - \left( \sum_{i=1}^n y_i \right)^2$$

$$SQ_{\text{Total}} = 8(4318) - (184)^2 = 688,00$$

Causas de Variação	G.L.	SQ	QM	F
Regressão	1	624,42	624,42000	58,93
Resíduo	6	63,58	10,59587	---
Total	7	688,00	---	---

$$H_0: \beta = 0 \text{ e } H_1: \beta \neq 0$$

$$\text{Comparando o } F_{\text{calc.}} = 58,93 \text{ com o } F_{\text{tab.}} = F_{0,05 (1,6)} = 5,99$$

Conclui-se que a regressão de  $y$  sobre  $x$  segundo o modelo  $\hat{y}_i = 0,0159x_i + 15,66$  é significativa ao nível de significância de 5%. Uma vez estabelecida e testada a equação de regressão, a mesma pode ser usada para explicar o relacionamento entre as variáveis e também para fazer previsões dos valores de  $Y$  para valores fixados de  $X$ .

## Saídas do Statistica

Analysis of Variance: DV: VENDAS (regress.sta)					
Continue...	Sums of Squares	df	Mean Squares	F	p-level
Regress.	280212,6	1	280212,6	59,29733	,000251
Residual	28353,3	6	4725,6		
Total	308565,9				

### 10.7 Coeficiente de Determinação ( $r^2$ )

É o grau em que as predições baseadas na equação de regressão superam as predições baseadas em  $\bar{y}$ , ou ainda é a proporção entre a variância explicada pela variância total.

Variância Total = soma dos desvios ao quadrado

$$VT = SQTotal = \sum_{i=1}^n (y_i - \bar{y})^2 = n \sum_{i=1}^n y_i^2 - \left( \sum_{i=1}^n y_i \right)^2,$$

Variância Não-explicada = soma de quadrados dos desvios em relação a reta  $\hat{y}$

$$V\tilde{N}E = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Para facilitar os cálculos usaremos:

$$r^2 = \frac{\left( n \sum_{i=1}^n xy - \sum_{i=1}^n x \sum_{i=1}^n y \right)^2}{\left[ n \sum_{i=1}^n x^2 - \left( \sum_{i=1}^n x \right)^2 \right] \left[ n \sum_{i=1}^n y^2 - \left( \sum_{i=1}^n y \right)^2 \right]} = \frac{COV_{xy}}{S_{xx} \cdot S_{yy}}$$

$$r^2 = \frac{[8(89802) - (3691)(184)]^2}{[8(2011501) - (3691)^2][8(4318) - (184)^2]} = 0.908$$

O valor de  $r^2$  varia de 0 a 1, logo o fato de  $r^2 = 0.908$  (no exemplo), indica que aproximadamente 91% da variação do lucro estão relacionados com a variação das vendas, em outras palavras 9% da variação dos lucros não são explicados pelas vendas.

## Saídas do Statistica

Regression Summary for Dependent Variable: VENDAS						
Continue...	R= .95294944 R <sup>2</sup> = .90811263 Adjusted R <sup>2</sup> = .89279807 F(1,6)=59,297 p<.00025 Std.Error of estimate: 68,743					
N=8	BETA	St. Err. of BETA	B	St. Err. of B	t(6)	p-level
Intercept			-851,497	172,2159	-4,94436	.002593
LUCRO	.952949	.123752	57,081	7,4127	7,70048	.000251

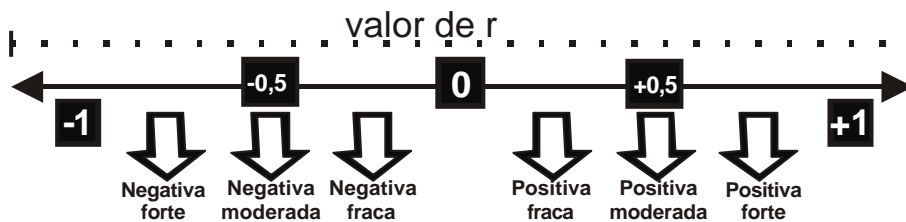
### 10.8 Coeficiente de Correlação (r)

Tem como objetivo encontrar o grau de relação entre duas variáveis, ou seja, um coeficiente de correlação.

Esta é a forma mais comum de análise, envolvendo dados contínuos, conhecido como "**r de Pearson**".

#### 10.8.1 Características do "r"

- Pode assumir valores positivos (+) como negativos (-), é semelhante ao coeficiente de regressão de uma reta ajustada num diagrama de dispersão;
- A magnitude de r indica quão próximos da "reta" estão os pontos individuais;
- quando o r se aproxima de +1 indica pouca dispersão, e uma correlação muito forte e positiva;
- quando o r se aproxima de "zero" indica muita dispersão, e uma ausência de relacionamento;
- quando o r se aproxima de -1 indica pouca dispersão, e uma correlação muito forte e negativa.





## 10.8.2 Medidas de Correlação

### 10.8.2.1 Tratamento Qualitativo (correlação momento produto)

Relação entre as variáveis, mediante a observação do diagrama de dispersão.

### 10.8.2.2 Tratamento Quantitativo

É o estabelecimento das medidas de correlação.

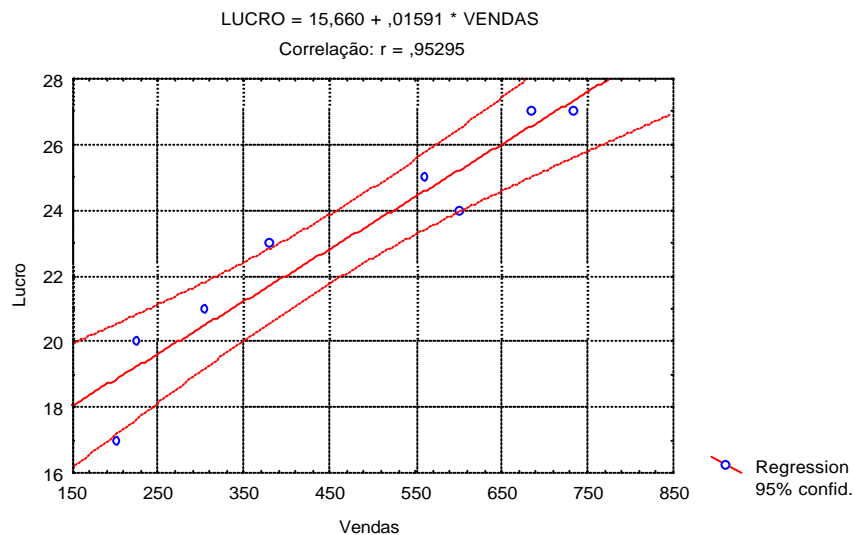
O valor de "r" pode ser enganoso, na realidade, uma estatística mais significativa é o  $r^2$  (coeficiente de determinação), que dá o valor percentual da variação de uma variável explicativa em relação a outra variável.

$$r = \frac{\left( n \sum_{i=1}^n xy - \sum_{i=1}^n x \sum_{i=1}^n y \right)}{\sqrt{\left[ n \sum_{i=1}^n x^2 - \left( \sum_{i=1}^n x \right)^2 \right]} \sqrt{\left[ n \sum_{i=1}^n y^2 - \left( \sum_{i=1}^n y \right)^2 \right]}} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

Logo podemos observar que o coeficiente de determinação nos dá uma base intuitiva para a análise de correlação.

No exemplo temos  $r = 0.953$

### Saída do Statistica



## 10.9 Exercícios

Para os dados abaixo:

- Construa um diagrama de dispersão;
- Determine a reta de regressão;
- Faça uma análise de variância do modelo;
- Calcule o Coeficiente de Explicação;
- Calcule o Coeficiente de Correlação de Pearson;
- Interprete os resultados obtidos.

X = 1º Exame e Y = 2º Exame

Aluno	1	2	3	4	5	6	7	8	9	10
Exame 1	82	84	86	83	88	87	85	83	86	85
Exame 2	92	91	90	92	87	86	89	90	92	90

$$R = \hat{y} = -0.79 x + 157.25; r^2 = 53.29\%; r = -0.73$$

X = horas de estudo e Y = Nota da Prova

Aluno	1	2	3	4	5	6	7	8
Horas	2	4	5	5	6	8	9	10
Nota	1	3	6	6	8	7	8	10

$$R = \hat{y} = 0.98 x + 0.12; r^2 = 82.99\%; r = +0.911$$

X = Seguro (x 1000 ) e Y = Renda (x100)

Indivíduo	1	2	3	4	5	6	7	8
Seguro	20	16	34	23	27	32	18	22
Renda	64	61	84	70	88	92	72	77

$$R = \hat{y} = 1.50 x + 40.08; r^2 = 74.3\%; r = +0.86$$

X = Peso do Pai (kg) e Y = Peso do Filho (kg)

Indivíduo	1	2	3	4	5	6	7	8	9	10
Peso Pai	65	63	67	64	68	62	70	66	68	67
Peso Filho	68	66	68	65	69	66	68	65	71	67

$$R = \hat{y} = 0.48 x + 35.48; r^2 = 40.58\%; r = +0.637$$

# 11 Referências Bibliográficas

---

COSTA NETO, P.L.O.; CYMBALISTA, M. (1994). **Probabilidades**. São Paulo: Edgard Blucher.

FONSECA, J.S.; MARTINS, G.A. (1993). **Curso de estatística** 4<sup>a</sup> ed. São Paulo: Atlas.

LAPONNI, Juan Carlos (1997). **Estatística usando o Excel** São Paulo: Laponni Treinamento e Editora.

MILONE, G.; ANGELINI, F. (1995). **Estatística aplicada** São Paulo: Atlas.

SNEDECOR, G. W.; COCHRAN, W. G. (1989). **Statistical Methods**. 8<sup>rd</sup> ed. Iowa: Iowa State University Press, 1989.

WONNACOTT, T.H.; WONNACOTT, R. J. (1990). **Introductory Statistics** New York. John Wiley & Sons;