UNIVERSIDADE ESTADUAL DE LONDRINA

CENTRO DE CIÊNCIAS EXATAS – CCE

DEPARTAMENTO DE ESTATÍSTICA

Curso de Especialização "Lato Sensu" em Estatística

ANÁLISE EXPLORATÓRIA DE DADOS

Professor: Dr. Waldir Medri

medri@uel.br

Londrina/Pr

Março de 2011

ÍNDICE

ESTATÍSTICA

1 INTRODUÇÃO

Desde a Antigüidade vários povos já registravam o número de habitantes, de nascimento, de óbitos, faziam estimativas das riquezas individual e social, distribuíam equitativamente terras ao povo, cobravam impostos e até realizavam inquéritos quantitativos por processos que, hoje, se chama de Estatística.

A palavra “Estatística” vem de status, que significa em latim Estado. Com essa palavra faziam-se as descrições e dados relativos aos Estados, tornando a Estatística um meio de administração para os governantes. Mais recentemente se passou a falar em estatística em várias ciências de todas as áreas do conhecimento humano, onde pode definir a Estatística como “um conjunto de métodos e processos quantitativos que servem para estudar e medir os fenômenos coletivos”.

Ao se estudar os fenômenos coletivos, o que interessa são os fatos que envolvem os elementos desses fenômenos, como eles se relacionam e qual o seu comportamento. Para que tal estudo possa acontecer com toda a seriedade que a ciência exige, é necessário que o levantamento seja feito através de uma pesquisa científica, sendo ela definida como a realização concreta de uma investigação planejada, desenvolvida e redigida de acordo com as normas de metodologia.

A Estatística é muito mais do que a simples construção de gráficos e o cálculo de médias. As informações numéricas são obtidas com a finalidade de acumular informação para a tomada de decisão. Então, a estatística pode ser vista como um conjunto de técnicas para planejar experimentos, obter dados e organizá-los, resumi-los, analisá-los, interpretá-los e deles extrair conclusões.

A informação de estatística é apresentada constantemente no rádio e na televisão, como por exemplo, a coleta de dados sobre nascimentos e mortes, a avaliação da eficiência de produtos comerciais e a previsão do tempo.

As técnicas clássicas da estatística foram delineadas para serem as melhores possíveis sob rigorosas suposições. Entretanto, a experiência tem forçado os estudiosos a conhecer que as técnicas clássicas comportam-se mal quando situações práticas não apresentam o ideal descrito por tais suposições. O

desenvolvimento recente de métodos exploratórios robustos está aumentando a eficiência da análise estatística.

Os bons profissionais de estatística têm sempre olhado com detalhes os dados antes de levantar suposições estatísticas e testes de hipóteses. Mas o uso indiscriminado de pacotes estatísticos computacionais, sem o exame cuidadoso dos dados profissionais da área, conduz, às vezes, a resultados aberrantes.

A análise exploratória de dados nos fornece um extenso repertório de métodos para um estudo detalhado dos dados, antes de adaptá-los. Nessa abordagem, a finalidade é obter dos dados a maior quantidade possível de informação, que indique modelos plausíveis a serem utilizados numa fase posterior, a *análise confirmatória de dados ou inferência estatística*.

2 ÁREAS DA ESTATÍSTICA

Se entender Estatística como a Ciência dos Dados, será de grande valia o domínio que seu corpo de conhecimento pode oferecer. Primeiramente, como ponto de partida, pode-se dividir a Estatística em duas áreas:

- Descritiva
- Inferencial (Indutiva)

Obs. Alguns autores, como por exemplo, Marcos Nascimento Magalhães e Antonio Carlos Pedroso de Lima, dizem que a estatística, grosso modo, pode ser dividida em três áreas: Estatística descritiva; Probabilidade e Inferência estatística.

2.1 ESTATÍSTICA DESCRITIVA

A Estatística Descritiva se preocupa com a organização, apresentação e sintetização de dados. Utilizam gráficos, tabelas e medidas descritivas como ferramentas. Utilizada na etapa inicial da análise, destinada a obter informações que indicam possíveis modelos a serem utilizados numa fase final que seria a chamada inferência estatística.

2.2 ESTATÍSTICA INFERENCIAL

A Estatística Inferencial postula um conjunto de técnicas que permitem utilizar dados oriundos de uma amostra para generalizações sobre a população. Constitui esse conjunto de técnicas: a determinação do número de observações (tamanho da amostra); o esquema de seleção das unidades observacionais; o cálculo das medidas estatísticas; a determinação da confiança nas estimativas; a significância dos testes estatísticos; a precisão das estimativas; dentre outras. Essa generalização é feita a partir do processo de estimação das medidas estatísticas que podem ser calculadas, porém não sem antes se antecipar um grau de certeza de que a amostra esteja fornecendo os dados que seriam de se esperar caso toda a população fosse estudada. Nesse caso, o ramo da matemática que será utilizado para se avaliar tal grau de certeza é a probabilidade. Com ela teremos condições de mensurar a fidedignidade de cada inferência feita com base na amostra.

Antes de começar a estudar os métodos estatísticos que permitirá analisar dados, sejam eles qualitativos ou quantitativos, é importante introduzir alguns conceitos preliminares a fim não apenas de dar nomes aos instrumentos, mas também adequar e equalizar a terminologia a ser utilizada ao longo do curso.

Na terminologia estatística, o grande conjunto de dados que contém a característica que temos interesse recebe o nome de *população.* Esse termo refere-se não somente a uma coleção de indivíduos, mas também ao alvo sobre o qual reside nosso interesse. Assim, nossa população pode ser tanto todos os habitantes de Londrina como todas as lâmpadas produzidas por uma fábrica em certo período de tempo. Algumas vezes podemos acessar toda a população para estudarmos características de interesse, mas, em muitas situações, tal procedimento não pode ser realizado. Em geral, razões econômicas são determinantes dessas situações. Por exemplo, uma empresa, usualmente, não dispõe de verba suficiente para saber o que pensam todos os consumidores de seus produtos. Há ainda razões éticas, quando, por exemplo, os experimentos de laboratório que envolvem o uso de seres vivos. Além disso, existem casos em que a impossibilidade de se acessar toda a população de interesse é incontornável. Por exemplo, em um experimento para determinar o tempo de funcionamento das lâmpadas produzidas por uma indústria, não podemos observar toda a população de interesse.

Tendo em vista as dificuldades de várias naturezas para se observar todos os elementos da população, tomaremos alguns deles para formar um grupo a ser estudado. Este subconjunto da população, em geral com dimensão menor, é denominado *amostra*.

3 POPULAÇÃO E AMOSTRA

3.1 POPULAÇÃO

População é o conjunto constituído por todos os indivíduos que representam pelo menos uma característica comum, cujo comportamento interessa analisar (inferir). Assim sendo, o objetivo das generalizações estatísticas está em dizer se algo acerca de diversas características da população estudada, com base em fatos conhecidos.

3.2 AMOSTRA

Amostra pode ser definida como um subconjunto, uma parte selecionada da totalidade de observações abrangidas pela população, através da qual se faz inferência sobre as características da população. Uma amostra tem que ser representativa, a tomada de uma amostra bem como seu manuseio requer cuidados especiais para que os resultados não sejam distorcidos.

- **Parâmetro** é uma medida numérica que descreve uma característica de uma população. São valores fixos, geralmente desconhecidos e usualmente representados por caracteres gregos. Por exemplo, μ (média populacional), p (proporção populacional), σ (desvio-padrão populacional), σ^2 (variância populacional).
- **Estatística** é uma estatística numérica que descreve uma característica de uma amostra. Representada por caracteres latinos. Por exemplo, $\bar{x}$ (média amostral), $\hat{p}$ (proporção amostral), s (desvio-padrão amostral), s^2 (variância amostral).
- **Unidade Observável** é a portadora da(s) característica(s), ou propriedade(s), que se deseja investigar.

A seleção da amostra pode ser feita de várias maneiras, dependendo, entre outros fatores, do grau de conhecimento que temos da população, da quantidade de recursos disponíveis a assim por diante. Cabe ressaltar que este item será apresentado mais para frente.

4 VARIÁVEIS

Ao se fazer um estudo estatístico de um determinado fato ou grupo, tem-se que considerar o tipo de variável. Pode ter *variáveis qualitativas* ou *variáveis quantitativas.*

4.1 VARIÁVEIS QUALITATIVAS

Variáveis qualitativas são aquelas em que a variável assume “valores” em categorias, classes ou rótulos. São, portanto, por natureza, dados não numéricos. Apesar de ser considerada de baixo nível de mensuração, do ponto de vista da aplicação de instrumental estatístico, a variável qualitativa oferece um vasto espectro de aplicação nas ciências sociais e do comportamento. Variáveis qualitativas denotam características individuais das unidades sob análise, tais como sexo, estado civil, naturalidade, raça, grau de instrução, dentre outras, permitindo estratificar as unidades para serem analisadas de acordo com outras variáveis.

4.2 VARIÁVEIS QUANTITATIVAS

Variáveis quantitativas são aquelas expressas pelas variáveis com níveis de mensuração intervalar ou de razão. Ou seja, são aqueles nas quais as variáveis assumem valores numa escala métrica definida por uma origem e uma unidade, por exemplo: idade, salário, peso, etc.

As variáveis qualitativas podem ser, também, classificadas como *nominal* e *ordinal*. Por outro lado, as variáveis quantitativas podem ser classificadas como *discretas*, quando assumem um número finito de valores, ou *contínuas*, quando assume um número infinito de valores, geralmente em intervalos, como apresentam na Tabela 1.

Tabela 1: Classificação das variáveis qualitativas e quantitativas

Variáveis	Tipos	Descrição	Exemplos
Qualitativas ou Categóricas	Nominal	Não existe nenhuma ordenação	Cor dos olhos, sexo, estado civil, tipo sangüíneo.
	Ordinal	Existe uma ordenação I, II, III	Nível de escolaridade, estágio da doença, colocação de concurso.
Quantitativas	Discretas	Valor pertence a um conjunto enumerável	Número de filhos por casal, quantidade de leitos
	Contínuas	Quando o valor pertence a um intervalo real	Medidas de altura e peso, taxa de glicose, nível de colesterol.

Em algumas situações podem-se atribuir valores numéricos às várias qualidades ou atributos e depois proceder à análise como esta variável como se fosse quantitativa, desde que o procedimento seja passível de interpretação.

Uma vez obtidos os dados referentes às variáveis qualitativas, a tarefa seguinte é representá-los através de uma tabela e de um gráfico. Posteriormente, poderá ser útil calcular as frequências, simples, acumuladas e as relativas.

Para os dados quantitativos, quando o número de observações cresce e os valores são diferenciados entre si, há que se representá-los de modo resumido. Para isso a melhor forma de representação tabular é através de distribuições de frequência por classes de valores.

Como exemplo: Suponha que um médico está interessado em fazer um levantamento sobre algumas características de pacientes atendidos em sua clínica neurológica: sexo peso, tipo de tratamento, número de convulsões e classificação da doença (leve, moderada e severa).

Os dados podem ser organizados em uma tabela. Usualmente os indivíduos são representados nas linhas e as variáveis nas colunas. Este formato é utilizado pela maioria do programas computacionais.

Note através da Tabela 2 que cada indivíduo é uma unidade de observação na qual são feitas várias medidas e/ou anotados vários atributos, referentes às variáveis.

Tabela 2: Características de pacientes atendidos em uma clínica neurológica

Paciente	Sexo	Peso	Tipo de Tratamento	Nº de Convulsões	Classificação da Doença
1	M	89,8	A	1	Leve
2	F	64,2	A	3	Severa
3	M	91,0	B	2	Moderada
4	F	56,7	A	0	Moderada
5	F	48,5	B	1	Leve
. . .					
58	M	71,0	B	0	Severa
59	M	78,8	A	2	Leve
60	F	71,0	B	3	Moderada

Analise a tabela 2 e classifique as variáveis:

- Variáveis qualitativas nominal: Sexo, Tipo de tratamento.
- Variáveis qualitativas ordinal: Classificação da doença.
- Variáveis quantitativas discreta: Número de convulsões
- Variáveis quantitativas contínua: Peso.

Um outro exemplo: Um pesquisador está interessado em fazer um levantamento sobre alguns aspectos socioeconômicos dos empregados da seção de orçamentos da Companhia MB. Usando informações obtidas do departamento pessoal, ele elaborou a Tabela 3.

De modo geral, para cada elemento investigado numa pesquisa, tem-se associado um (ou mais de um) resultado correspondendo à realização de uma característica (ou características).

Algumas variáveis, como sexo, educação, estado civil, apresentam como possíveis realizações de qualidade (ou atributo) do indivíduo pesquisado, ao passo que outras, como número de filhos, salário, idade, apresentam como possíveis realizações números resultantes de uma contagem ou mensuração. As variáveis do primeiro tipo são chamadas *qualitativas* e as do segundo *quantitativas*.

Tabela 3: Informações sobre estado civil, grau de instrução, número de filhos, salário mínimo, idade e procedência de 36 empregados da seção de orçamentos da companhia MB.

Nº	Estado Civil	Grau de Instrução	Nº de Filhos	Salário mínimo	Idade Anos	Meses	Região de Procedência
1	Solteiro	Ensino fundamental		4,00	26	3	Interior
2	Casado	Ensino fundamental	1	4,56	32	10	Capital
3	Casado	Ensino fundamental	2	5,25	36	5	Capital
4	Solteiro	Ensino médio		5,73	20	10	Outra
5	Solteiro	Ensino fundamental		6,26	40	7	Outra
6	Casado	Ensino fundamental	0	6,66	28	0	Interior
7	Solteiro	Ensino fundamental		6,86	41	0	Interior
8	Solteiro	Ensino fundamental		7,39	43	4	Capital
9	Casado	Ensino médio	1	7,44	34	10	Capital
10	Solteiro	Ensino médio		7,59	23	6	Outra
11	Casado	Ensino médio	2	8,12	33	6	Interior
12	Solteiro	Ensino fundamental		8,46	27	11	Capital
13	Solteiro	Ensino médio		8,74	37	5	Outra
14	Casado	Ensino fundamental	3	8,95	44	2	Outra
15	Casado	Ensino médio	0	9,13	30	5	Interior
16	Solteiro	Ensino médio		9,35	38	8	Outra
17	Casado	Ensino médio	1	9,77	31	7	Capital
18	Casado	Ensino fundamental	2	9,80	39	7	Outra
19	Solteiro	Ensino superior		10,35	25	8	Interior
20	Solteiro	Ensino médio		10,76	37	4	Interior
21	Casado	Ensino médio	1	11,06	30	9	Outra
22	Solteiro	Ensino médio		11,59	34	2	Capital
23	Solteiro	Ensino fundamental		12,00	41	0	Outra
24	Casado	Ensino superior	0	12,79	26	1	Outra
25	Casado	Ensino médio	2	13,23	32	5	Interior
26	Casado	Ensino médio	2	13,60	35	0	Outra
27	Solteiro	Ensino fundamental		13,85	46	7	Outra
28	Casado	Ensino médio	0	14,69	29	8	Interior
29	Casado	Ensino médio	5	14,71	40	6	Interior
30	Casado	Ensino médio	2	15,99	35	10	Capital
31	Solteiro	Ensino superior		16,22	31	5	Outra
32	Casado	Ensino médio	1	16,61	36	4	Interior
33	Casado	Ensino superior	3	17,26	43	7	Capital
34	Solteiro	Ensino superior		18,75	33	7	Capital
35	Casado	Ensino médio	2	19,40	48	11	Capital
36	Casado	Ensino superior	3	23,30	42	2	Interior

Fonte: Dados hipotéticos

5 DADOS

São as informações inerentes às variáveis que caracterizam os elementos que constituem a população ou a amostra em estudo. Os dados obtidos em pesquisas devem ser analisados e interpretados com o auxílio de métodos estatísticos.

Na primeira etapa deve-se fazer uma análise descritiva que consiste na organização e descrição dos dados, na identificação de valores que representem o elemento típico e, na quantificação da variabilidade presente nos dados.

5.1 DADOS BRUTOS

Qualquer pesquisa é baseada em levantamento ou coleta de dados. Os dados são obtidos diretamente da pesquisa, sem terem passados por nenhum processo de síntese ou análise. Por exemplo, os 50 valores, em decibéis, de nível de ruído de tráfego em certo cruzamento estão apresentados a seguir:

58,0	62,5	65,0	67,0	68,3	65,0	66,4	58,0	67,0	67,0
62,5	62,5	66,4	66,4	65,0	65,0	60,2	60,2	60,2	60,2
59,5	59,5	59,5	65,0	66,4	66,4	66,4	60,2	62,5	67,0
67,0	67,0	70,1	70,1	71,9	70,1	67,0	66,4	66,4	68,3
68,3	68,3	65,0	65,0	62,5	62,5	65,0	65,0	68,3	71,9

Apesar de todos estes valores terem sido obtidos em de nível de ruído de tráfego em certo cruzamento, nota-se uma grande variação nos resultados. Assim, os métodos estatísticos são fundamentais para o estudo de situações em que a variabilidade é inerente. A *Estatística Descritiva* ajuda na percepção, avaliação e quantificação da variabilidade em tabelas e gráficos obtidos a partir de um conjunto de dados que sintetizem os valores, com o objetivo de se ter uma visão global e clara da variação existente nas variáveis.

5.2 ROL

A mão, ou com auxílio de computador, pode-se classificar os dados x_1, x_2,...,x_n em ordem crescente. Pode-se, pelo rol, verificar de maneira mais clara e rápida a composição do conjunto, identificando o maior e o menor valor além de alguns elementos que podem se repetir várias vezes, mostrando assim o comportamento dos dados.

5.3 DISPOSITIVO - RAMO E FOLHAS

A mais comum estrutura de dados é um grupo de números. Até mesmo esta tão simples estrutura de dados pode ter características não facilmente distinguíveis por estudos dos números. *O dispositivo "ramo e folhas"* é uma técnica flexível e eficaz para começarmos a olhar um conjunto ou uma amostra de dados. Os dígitos mais significantes dos valores, por si próprios, fazem muito trabalho de ordenação do grupo.

Está técnica básica, mas versátil, é intensamente usada, principalmente para comparar grupos e examinar cada característica, tais como:

- quanto o grupo está próxima da assimetria;
- como estão distribuídos os valores;
- se alguns valores estão distanciados dos demais;
- se existe concentração de dados;
- se existe lacunas nos dados.

Aplicação do dispositivo "ramo e folhas". Não existe uma regra fixa para construir o ramo e folhas, mas a idéia básica é dividir cada observação em duas partes: a primeira (o ramo) é colocada à esquerda de uma linha vertical, a segunda (a folha) é colocada à direita. A Figura 1 apresenta um dessa aplicação.

Ramo	Folha	Frequência
58	0 0	2
59	5 5 5	3
60	2 2 2 2 2	5
62	5 5 5 5 5 5	6
65	0 0 0 0 0 0 0 0 0	9
66	4 4 4 4 4 4 4 4	8
67	0 0 0 0 0 0 0	7
68	3 3 3 3 3	5
70	1 1 1	3
71	9 9	2

Figura 1 - Ramos e folhas para os depósitos bancários

Assim, o Rol dos 50 valores do nível de ruído de tráfego em certo cruzamento, faca:

58,0	58,0	59,5	59,5	59,5	60,2	60,2	60,2	60,2	60,2
62,5	62,5	62,5	62,5	62,5	62,5	65,0	65,0	65,0	65,0
65,0	65,0	65,0	65,0	65,0	66,4	66,4	66,4	66,4	66,4
66,4	66,4	66,4	67,0	67,0	67,0	67,0	67,0	67,0	67,0
68,3	68,3	68,3	68,3	68,3	70,1	70,1	70,1	71,9	71,9

A apresentação dos dados pode ser de duas formas: Apresentação Tabular e apresentação Gráfica.

5.4 REPRESENTAÇÃO TABULAR

Apresentação tabular numérica de dados é a representação das informações por intermédio de uma tabela. Uma tabela é uma maneira bastante eficiente de mostrar os dados levantados e que facilita a compreensão e interpretação dos dados.

Para organizar uma série estatística ou uma distribuição de frequências, existem algumas normas nacionais ditadas pela Associação Brasileira de Normas Técnicas (ABNT) as quais devem ser respeitadas. Assim, toda tabela estatística de conter:

a) Elementos essenciais

- **Título –** indica a natureza do fato estudado (o quê?), as variáveis escolhidas na análise do fato (como?), o local (onde?) e a época (quando?).
- **Corpo –** é o conjunto de linhas e colunas que contém, respectivamente, as séries horizontais e verticais de informações.
- **Cabeçalho –** designa a natureza do conteúdo de cada coluna.
- **Coluna indicadora –** mostra a natureza do conteúdo de cada linha.

b) Elementos complementares (se necessário)

- **Fonte –** é o indicativo, no rodapé da tabela, da entidade responsável pela sua organização ou fornecedora dos dados primários.
- **Notas –** são colocadas no rodapé da tabela para esclarecimentos de ordem geral.

c) Sinais convencionais

- **– (hífen)**, quando o valor numérico é nulo;
- **... (reticência)**, quando não se dispõe de dado;

- **? (ponto de interrogação)**, quando há dúvidas quanto à exatidão do valor numérico;
- **0; 0,0; 0,00 (zero)**, quando o valor numérico é muito pequeno para ser expresso pela unidade utilizada, respeitando o número de casas decimais adotado;
- **X (letra x)**, quando o dado for omitido.

d) Numerar as tabelas quando houver mais de uma.

e) As tabelas devem ser fechadas acima e abaixo por linha horizontal, não sendo fechadas à direita e à esquerda por linhas verticais. É facultativo o emprego de traços verticais para separação de colunas no corpo da tabela.

f) Os totais e subtotais devem ser destacados.

g) Manter a uniformidade do número de casas decimais.

As tabelas podem ser classificadas como unidimensional ou bidimensional. A Tabela 4 é uma representação unidimensional, enquanto a Tabela 5 é bidimensional.

Tabela 4: Número e porcentagem de causas de morte de residentes de Londrina, no período de 10 de agosto a 31 de dezembro de 2008

CAUSAS DA MORTE	**Nº**	**%**
Doenças do ap. circulatório	281	33,5
Neoplasias	115	13,7
Causas externas	92	11,0
Doenças do ap. respiratório	87	10,4
Doenças das glând. endóc./transt. Imunitários	56	6,7
Doenças do ap. digestivo	54	6,4
Doenças e infec. e parasitárias	46	5,5
Afecções do per. Perinatal	26	3,1
Demais grupos	82	9,8
TOTAL	**839**	**100,0**

FONTE: Núcleo de informação em mortalidade – PML

Tabela 5: Percentual de vendas do produto A, da Empresa WD, no mês de março de 2008

REGIÃO	FAIXA ETÁRIA				
	< 1 ano	1 a 4 anos	5 a 19 anos	20 a 49 anos	50 anos ou +
Centro	4,54	-	2,02	14,65	78,79
Norte	6,45	1,61	2,42	26,61	62,91
Sul	7,27	4,55	5,45	22,73	60,00
Leste	3,36	-	4,03	24,16	68,45
Oeste	4,57	1,14	3,43	18,29	72,57
Rural	15,71	4,29	4,28	14,29	61,43
LONDRINA	5,83	1,42	3,37	20,61	68,77

FONTE: Relatório do mês de março do Departamento de vendas.

5.5 Representação Gráfica

A representação gráfica é usada para aumentar a legibilidade do resultado de uma pesquisa. Os gráficos devem ser auto-explicativos e de fácil compreensão.

Devem sempre ter um título, onde se destaca o fato, o local e o tempo. Ser construídos em uma escala que não desfigure os fatos ou as relações que se deseja destacar. Assim, a altura de um gráfico deve compreender entre 60% a 80% da largura.

5.5.1 Representação Gráfica para uma Variável Qualitativa

Para esse tipo de variável os gráficos mais utilizados são os de: colunas, barras, linhas e de setores.

Tabela 6: Densidade demográfica, segundo as Grandes Regiões - 2008

Brasil e Grandes Regiões	Densidade demográfica (hab/km^2)
Brasil	22,3
Norte	4,0
Nordeste	34,4
Sudeste	86,3
Sul	47,8
Centro Oeste	8,6

Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílio 2008

No Brasil a densidade demográfica média, em 2008, é de 22,3 hab/km^2. Região Norte, que possui 45,2% da área total do País e 8,1% da população, tem apenas 4,0 hab/km^2 Nessa região, ainda existem grandes vazios espaciais, em função da vastidão territorial e de grandes áreas intocadas, como a ocupada pela floresta Amazônica. A Região Sudeste, a mais evoluída economicamente do País, com 42% da população total, é a que tem a maior densidade com 86,3 hab/km^2 A Região Metropolitana de São Paulo, com 19,5 milhões de pessoas, corresponde a 47,9% da população do estado, enquanto a Região Metropolitana do Rio de Janeiro, com 11,5 milhões de pessoas, contém 73,4% dos habitantes do Rio de Janeiro (Tabela 6).

a) Gráfico de Colunas

Os gráficos de colunas (Figura 2) ou barras (Figura 3) consistem em construir retângulos, em que uma das dimensões é proporciona à magnitude a ser representada, sendo a outra arbitrária, porém igual para todas as colunas (ou barras). Essas colunas (ou barras) são dispostas paralelamente umas às outras, verticalmente (ou horizontalmente), isto é:

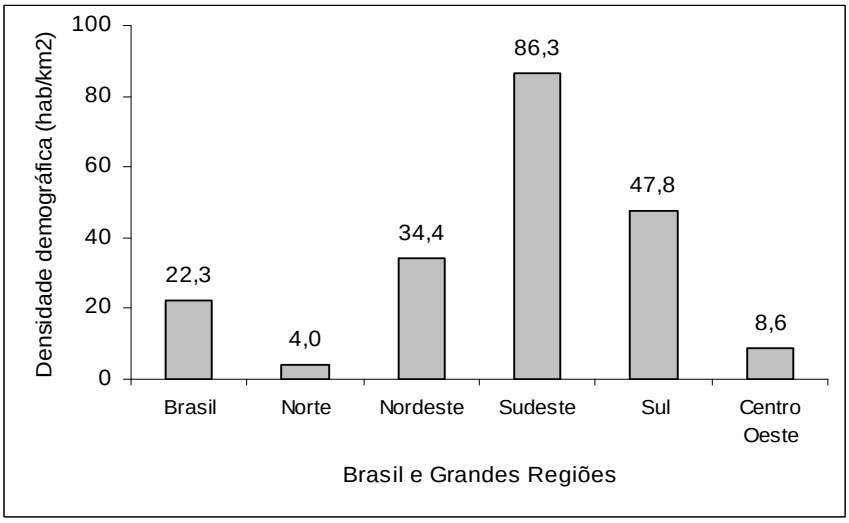


Figura 2 – Densidade demográfica, Brasil e as Grandes Regiões - 2008

b) Gráfico de Barras

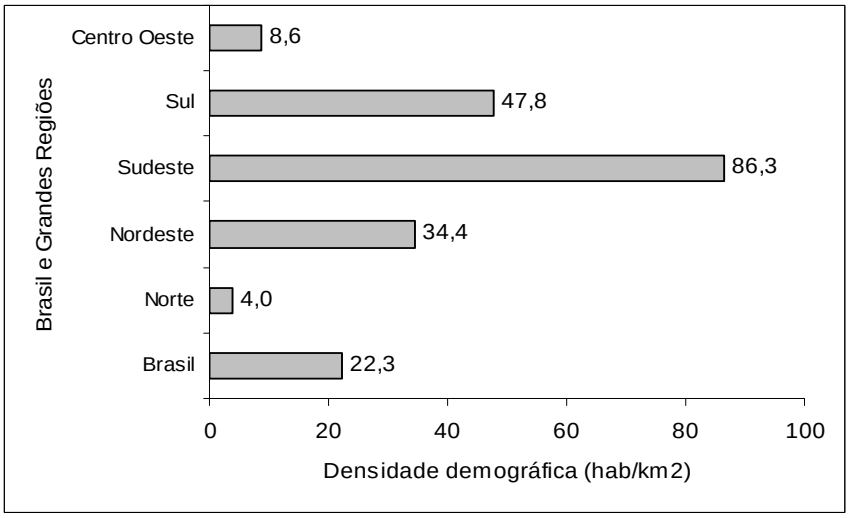


Figura 3 – Densidade demográfica, Brasil e as Grandes Regiões - 2008

c) Gráfico de Linhas (Figura 4)

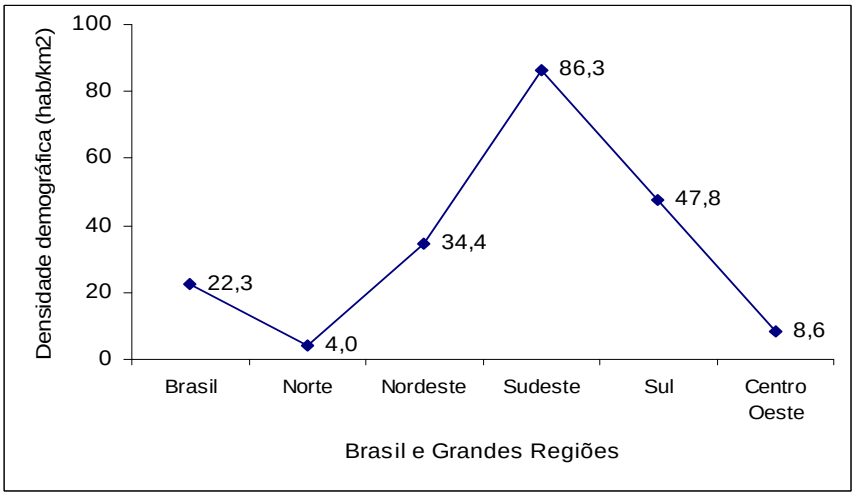


Figura 4 – Densidade demográfica, Brasil e as Grandes Regiões, 2008

Obs. *O gráfico de linha acima não é adequado para o exemplo*

d) Gráfico de Setores

O gráfico de setores (Figura 5) destina-se representar a composição, usualmente em porcentagem, de partes de um todo. Consiste num círculo de raio arbitrário, representando o todo, dividindo em setores, que correspondem às partes de maneira proporcional.

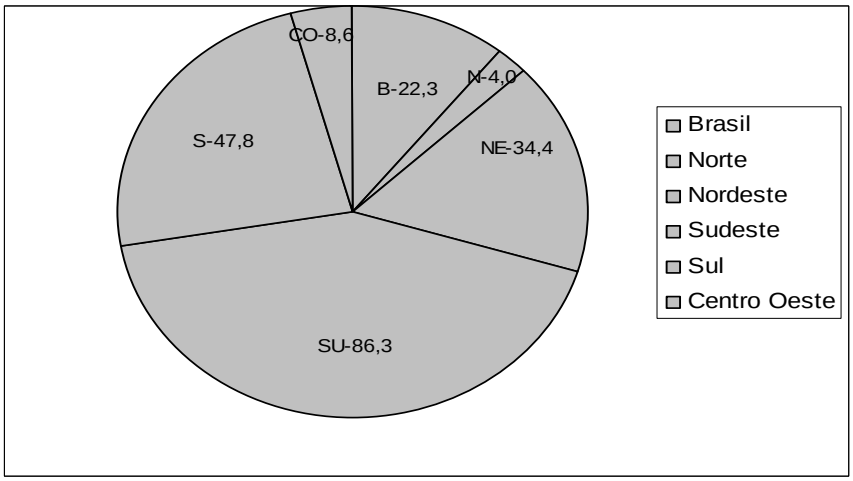


Figura 5 – Densidade demográfica, Brasil e as Grandes Regiões - 2008

5.5.2 Representação Gráfica para uma Variável Quantitativa

Gráficos referentes a variáveis quantitativas (discretas ou contínuas) mais utilizados são os de: colunas (Figura 6) e barras (Figura 7).

Tabela 7: As taxas mensais, em porcentagem, da Poupança, no período de janeiro a dezembro de 2005

Meses	Taxa (%)
Janeiro	0,715
Fevereiro	0,692
Março	0,675
Abril	0,734
Maio	0,737
Junho	0,739
Julho	0,774
Agosto	0,808
Setembro	0,771
Outubro	0,733
Novembro	0,711
Dezembro	0,714

Fonte: Caixa Econômica Federal

a) Gráfico de colunas

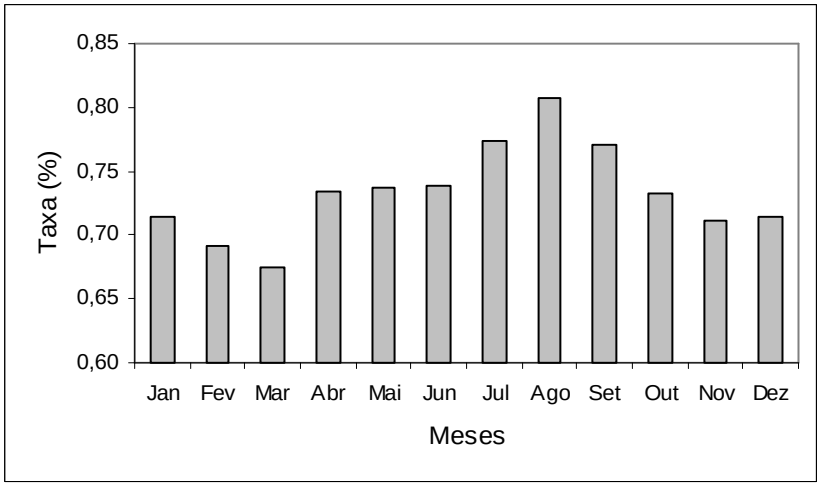


Figura 6 – Taxa de juros em porcentagem da caderneta de Poupança de janeiro a dezembro de 2005

c) Gráfico de linhas

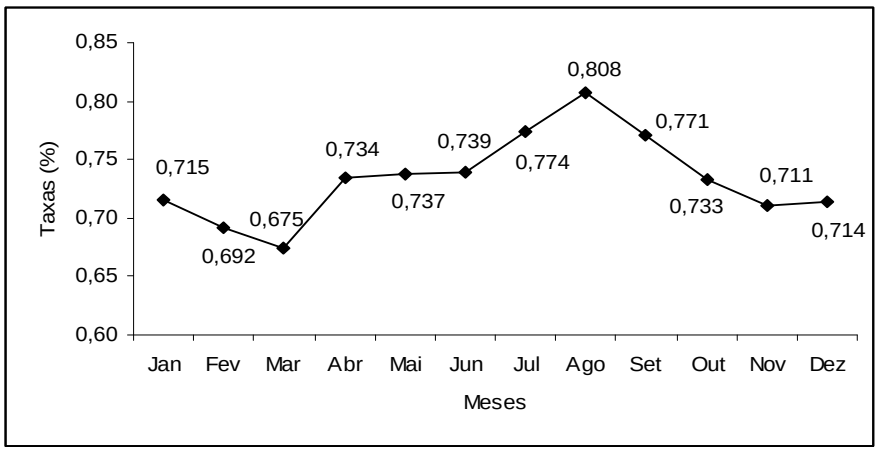


Figura 7 – Taxa de juros em porcentagem da caderneta de Poupança de janeiro a dezembro de 2005

5.5.3 Séries Conjugadas

Muitas vezes tem-se a necessidade de apresentar, em uma única tabela, a variação de valores de mais de uma variável, isto é, fazer uma conjunção de duas ou mais séries. Conjugando duas séries em uma única tabela, obtém-se uma tabela de dupla entrada (horizontal e vertical). A Tabela 8 apresenta a média de anos de estudo, no Brasil e nas Regiões: Sudeste e Nordeste, no período de 2002 a 2008

Tabela 8: Média de anos de estudo, no Brasil e nas Regiões, Sudeste e Nordeste, no período de 2002 a 2008

Brasil e Regiões	Anos						
	2002	2003	2004	2005	2006	2007	2008
Sudeste	7,2	7,4	7,6	7,7	7,9	7,9	8,1
Brasil	6,5	6,7	6,8	7,0	7,2	7,3	7,4
Nordeste	5,1	5,3	5,5	5,6	5,8	6,0	6,2

Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílio 2008

A educação básica no País é formada por dois ciclos – fundamental e médio – que correspondem a 11 anos de estudo completos. Os dados sobre os níveis de escolarização da população revelam melhoras, se comparados àqueles da década anterior, porém são ainda insuficientes e não compatíveis com o nível de desenvolvimento econômico do País. Basta observar a escolaridade média da população. Em 2008, o brasileiro de 15 anos ou mais de idade tinha, em média, 7,4 anos de estudo. Na Região Sudeste, essa média atingiu 8,1 anos, enquanto na Região Nordeste apenas 6,2 anos. Os com os gráficos, de linhas (figura 8) e de colunas múltiplas (figura 9) mostram esta situação.

a) Gráfico de Linhas (Figura 8)

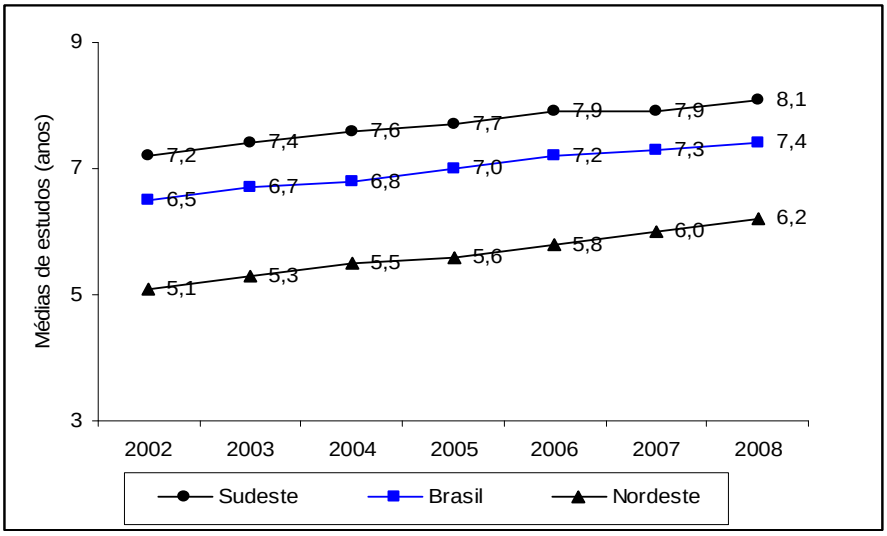


Figura 8 – Médias de estudo no Brasil e nas Regiões: Sudeste e Nordeste, no período de 2002 a 2008

b) Gráfico de Colunas Múltiplas (Figura 9)

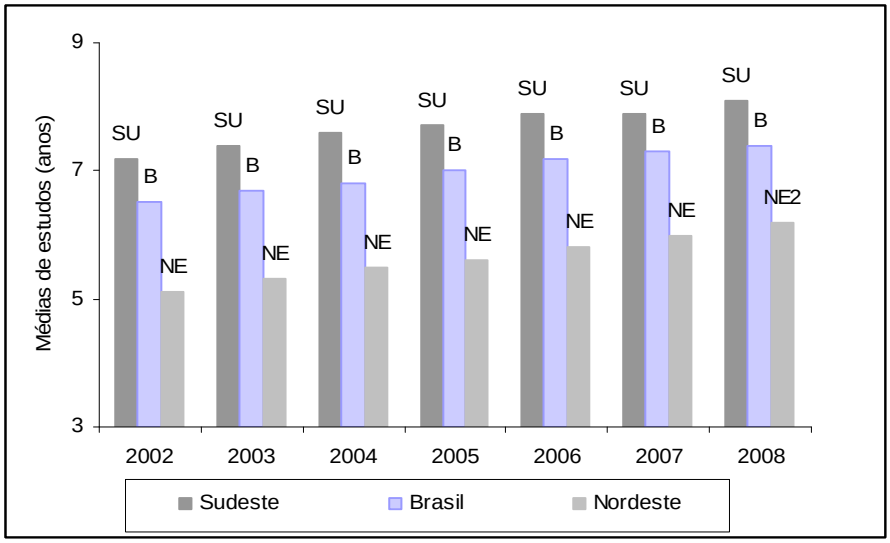


Figura 9 – Médias de estudo no Brasil e nas Regiões: Sudeste e Nordeste, no período de 2002 a 2008

O gráfico de colunas múltiplas é útil quando se quer fazer estudo comparativo.

5.5.4 Distribuição de Frequências

Quando se estuda uma variável, o maior interesse do pesquisador é conhecer o comportamento dessa variável, analisando a ocorrência de suas possíveis realizações. Considerando-se a variável qualitativa a ser estudada, como por exemplo, *grau de instrução* (Tabela 3), será observada e estudada muito mais facilmente quando se dispõem os ensinos: *Fundamental, Médio e Superior* em uma coluna e coloca-se, ao lado de cada ensino, o número de vezes que aparece repetido. Assim, a Tabela 9 apresenta a *distribuição de frequências* da variável grau de instrução.

Tabela 9: Frequências e porcentagens dos 36 empregados da seção de orçamentos da Companhia MB segundo o grau de instrução

Grau de Instrução	Frequência (n_i)	Proporção (f_i)	Porcentagem (%)
Fundamental	12	0,3333	33,33
Médio	18	0,5000	50,00
Superior	6	0,1667	16,67
Total	36	1,0000	100,00

Fonte: Tabela 3

Através da Tabela 9 da segunda coluna, nota-se que dos 36 empregados da Companhia MB, 12 têm o ensino fundamental, 18 o ensino médio e 6 possui curso superior.

Uma medida bastante útil na interpretação de tabelas de frequências é a proporção (ou a porcentagem) de cada realização em relação ao total. Assim 6/36 = 0,1667 (16,67%) dos empregados da Companhia MB (seção de orçamento) têm instrução superior. As proporções são muito úteis quando se quer comparar resultados de duas pesquisas distintas. Por exemplo, suponha-se que se queira comparar a variável *grau de instrução* para os empregados da seção de orçamentos com a mesma variável para todos os empregados da Companhia MB. Supondo que a empresa tenha 2.000 empregados e que a distribuição de frequências seja a Tabela 10.

Tabela 10: Frequências e porcentagens dos 2.000 empregados da Companhia MB segundo o grau de instrução

Grau de Instrução	**Frequência (n_i)**	**Proporção (f_i)**	**Porcentagem (%)**
Fundamental	650	0,3250	32,50
Médio	1.020	0,5100	51,00
Superior	330	0,1650	16,50
Total	2.000	1,0000	100,00

Fonte: dados hipotéticos

Importante: Não pode comparar diretamente as colunas das frequências das Tabelas 9 e 10, pois os totais de empregados são diferentes nos dois casos. Mas as colunas das porcentagens são comparáveis, já que as frequências foram reduzidas a um mesmo total. (no caso 100).

Gráficos para variáveis qualitativas

O gráfico de colunas múltiplas (Figura 10) segundo a variável qualitativa, *grau de instrução* das Tabelas 9 e 10, fica:

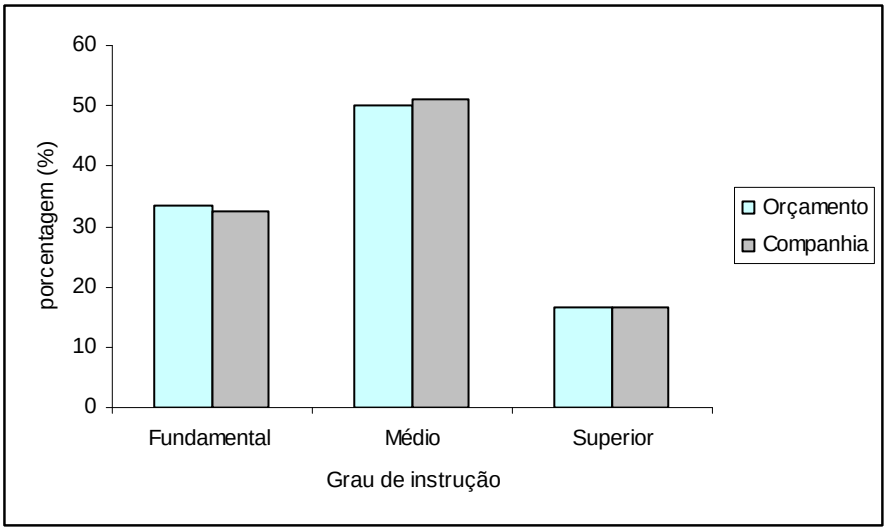


Figura 10 – Grau de instrução dos funcionários da Seção de Orçamento e da Companhia MB

Já o gráfico de linhas (Figura 11) referente a variável, *grau de instrução* das Tabelas 9 e 10, fica:

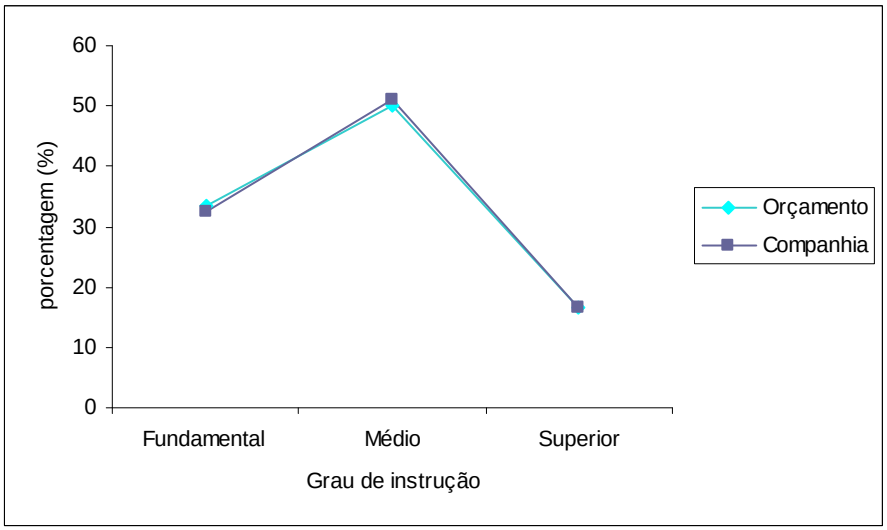


Figura 11 – Grau de instrução dos funcionários da Seção de Orçamento e da Companhia MB

Gráficos para variáveis quantitativas

Considerando-se, agora, a *variável quantitativa discreta* a ser estudada, *número de filhos* dos empregados casados da seção de orçamentos da Companhia MB (Tabela 3). A Tabela 11 apresenta a *distribuição de frequências* e as porcentagens desta variável.

Tabela 11: Frequências e porcentagens dos empregados da seção de orçamentos da Companhia MB, segundo o número de filhos

Nº de Filhos	**Frequência (n_i)**	**Porcentagem (%)**
0	4	20
1	5	25
2	7	35
3	3	15
5	1	5
Total	20	100

Fonte: Tabela 3

O gráfico de colunas (Figura 12) da variável quantitativa do *número de filhos* dos empregados casados da seção de orçamentos da Companhia MB da Tabela 11, é representado da seguinte forma:

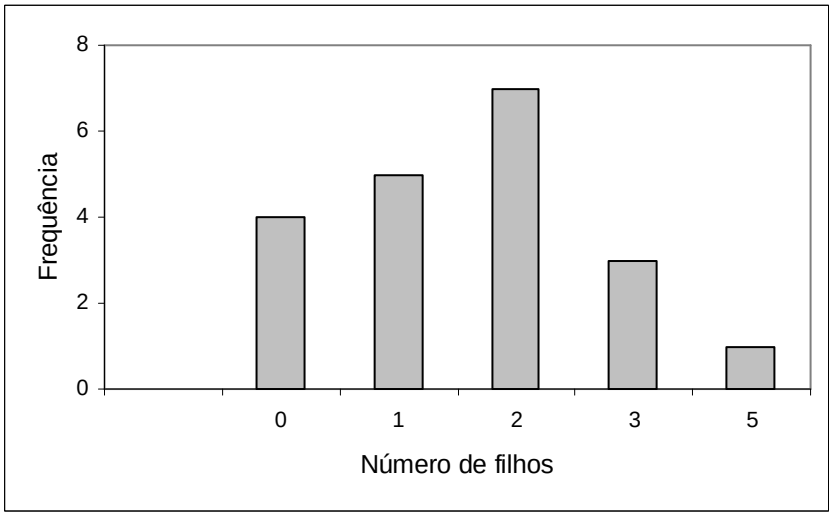


Figura 12 – Número de filhos dos empregados dos casados da seção de orçamento da Companhia

A construção de tabelas de frequências para *variáveis contínuas* necessita de certo cuidado. Por exemplo, a construção da tabela de frequências para a variável salário (Tabela 3) usando o mesmo procedimento anterior, não resumirá as 36 observações num grupo menor, pois não existem observações iguais. A solução empregada é agrupar os dados por faixas de salário. A Tabela 12 dá a distribuição de frequências dos salários dos 36 empregados da seção de orçamentos da Companhia MB por faixa de salários.

Tabela 12: Frequências e porcentagens dos 36 empregados da seção de orçamentos da Companhia MB por faixa de salário

Classe de Salários	**Frequência (n_i)**	**Porcentagem (%)**
4,00 |--- 8,00	10	27,78
8,00 |--- 12,00	12	33,33
12,00 |--- 16,00	8	22,22
16,00 |--- 20,00	5	13,89
20,00 |--- 24,00	1	2,78
Total	36	100,00

Fonte: Tabela 3

Procedendo-se desse modo, ao resumir os dados referentes a uma variável contínua, perde-se alguma informação. Por exemplo, não se sabe quais são os oito salários da classe de 12 a 16, a não ser que se investiga a tabela original (tabela 3). Sem perda de muita precisão, pode-se supor que todos os oito salários daquela classe fossem iguais ao ponto médio da referida classe, isto é, 14.

A distribuição de frequências é importante quando existe uma grande quantidade de dados. A finalidade em agrupar os dados é facilitar a visualização e também os cálculos deles, porém, a **determinação das medidas de posição e de dispersão** para uma variável quantitativa contínua, através de sua distribuição de frequências, exige aproximações, já que perde a informação dos valores observados.

Não há um modo único par se construir uma **tabela de frequência por classe de valores**. A escolha dos intervalos é arbitrária e a familiaridade do pesquisador com os dados é que lhe indicará quantas classes (intervalos) devem ser usadas. Entretanto, deve-se observar que, com um pequeno número de classes, perde-se informação, e com um número grande de classes, o objetivo de resumir os dados fica prejudicado. Estes dois extremos têm a ver, também, com o grau de suavidade da representação gráfica dos dados. Normalmente, sugere-se o uso de 5 a 15 classes com a mesma amplitude.

As classes não precisam ter amplitude constante, mas por uma questão de simplificação da construção da representação gráfica, geralmente são classes com

intervalos constantes. Por outro lado, existem técnicas para construção de tabelas de distribuição de frequências para intervalos contínuos (dados agrupados).

Etapas para a construção de tabelas de frequência para dados agrupados:

1) O cálculo da amplitude total dos dados é a diferença entre o maior e o menor valor da série, isto é:

   At = nº do maior – nº do menor

2) Não existindo um critério rígido para estabelecer o número ideal de intervalos, sugere-se que não se utilize menos de 5 e não mais de 15 intervalos. A experiência tem demonstrado que se pode fixar o número de intervalo como:

   $K = \sqrt{n}$ ou $K = 1 + 3,3.\log n$, para uma amostra de tamanho n

3) O intervalo das classes (amplitude de classes) pode ser feito dividindo-se a amplitude total pelo número de classes, isto é:

$$a_C = \frac{At}{K}$$

Assim, pode construir os intervalos partindo do menor valor do conjunto e somando a amplitude calculada (a_C), o que permite determinar os limites dos intervalos.

Aplicação: A Tabela 13 apresenta uma distribuição de frequência usando as técnicas de construção dos 50 valores, em decibéis, de nível de ruído de tráfego em certo cruzamento estão apresentados a seguir:

Cálculo:

$$\text{At} = X_{max} - X_{min} = 71,9 - 58,0 = 13,9$$

$$k = \sqrt{n} = \sqrt{50} \cong 7$$

$$a_C = \frac{At}{K} = \frac{13,9}{7} = \cong 2$$

Tabela 13: Nível de ruído, em decibéis, de tráfego em certo cruzamento

Nível de ruído (em db)	**Quantidade** (f_i)	**Ponto médio** ($\bar{x}_i$)	**Freq. Acum.** (F_{ac})	($x_i . f_i$)	($x_i^2 . f_i$)
58,0 |-- 60,0	5	59	5	295	17.405
60,0 |-- 62,0	5	61	10	305	18.605
62,0 |-- 64,0	6	63	16	378	23.814
64,0 |-- 66,0	9	65	25	585	38.025
66,0 |-- 68,0	15	67	40	1.005	67.335
68,0 |-- 70,0	5	69	45	345	23.805
70,0 |-- 72,0	5	71	50	355	25.205
Total	**50**			**3.268**	**214.194**

Os resultados referentes a variáveis contínuas frequentemente são organizados em tabelas de distribuições de frequências por intervalos. Três tipos de gráficos geralmente são utilizados neste caso: *histograma, polígono de frequência* e *ogivas.*

a) Histograma (Figura 13) é a representação gráfica de uma distribuição de frequência por meio de retângulos justapostos, contendo as classes de valores na abscissa e as frequências, absolutas ou relativas, nas ordenadas, centradas nos pontos médios.

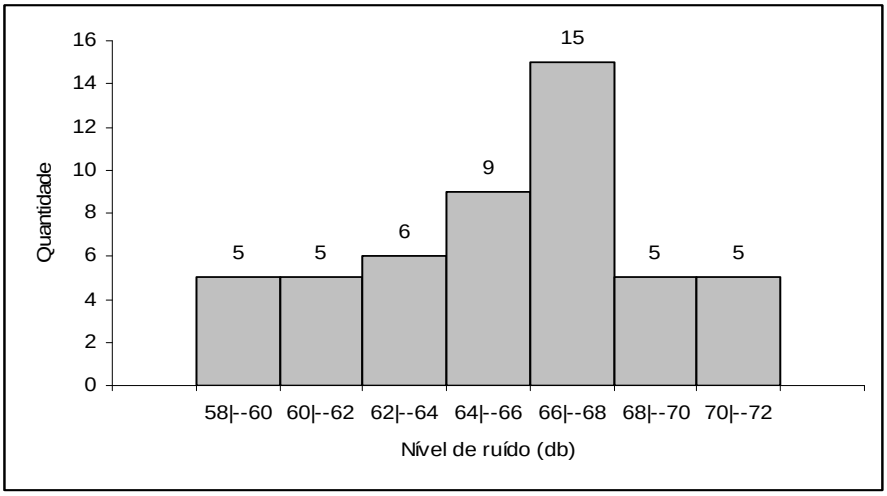


Figura 13 – Nível de ruído (db) em certo cruzamento

Através da figura, pode-se dizer que 10 níveis de ruído foram inferiores a 62 decibéis, ou 5 níveis de ruído foram iguais ou superiores a 70 decibéis.

b) **Polígono de frequências** (Figura 14) é a representação gráfica de uma distribuição de frequência, contendo os pontos médios de cada classe na abscissa e as frequências, absolutas ou relativas, nas ordenadas.

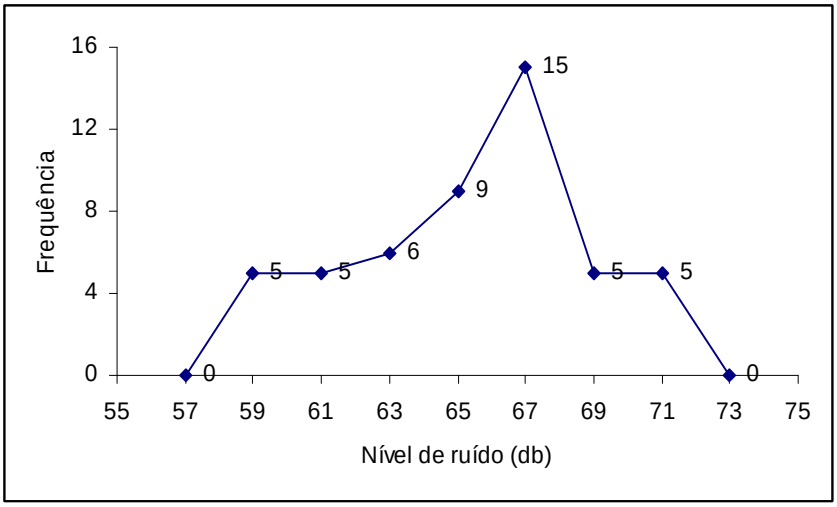


Figura 14 – Nível de ruído (db) em certo cruzamento

O gráfico de uma distribuição cumulativo é chamado de ogiva (Figura 15). Os valores dos dados são mostrados no eixo horizontal e as frequências cumulativas são apresentadas no eixo vertical.

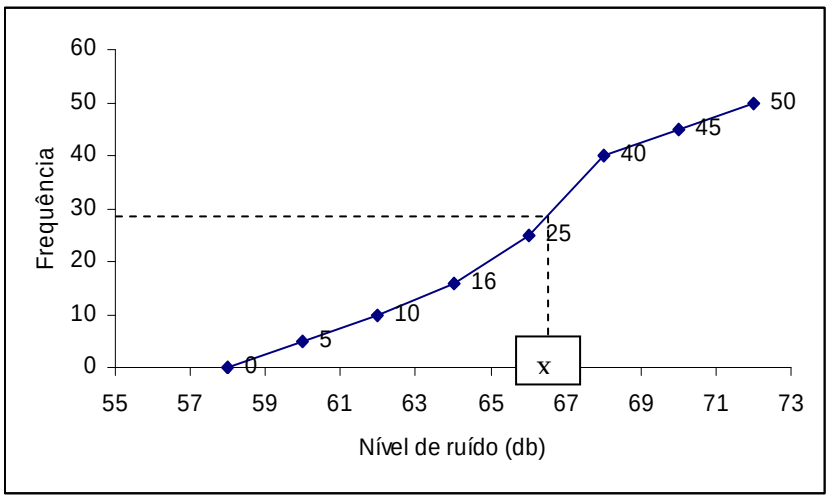


Figura 15 – Nível de ruído (db) acumulado em certo cruzamento

As frequências nesse exemplo foram acumuladas de modo crescente. Há casos, no entanto, que a acumulação das frequências é feita de modo decrescente. Este gráfico pose ser usado para fornecer informações adicionais. Por exemplo, para saber qual o nível de ruído **x** tal que 30 das quantidades (frequências) atingem menos do que x, basta procurar o ponto (x, 30) na curva. Observando as linhas pontilhadas no gráfico, nota-se que a solução é aproximadamente 67 decibéis.

5.6 LISTA 1 – EXERCÍCIOS

1) Ao nascer, os bebês são pesados e medidos, para se saber se estão dentro das tabelas de peso e altura esperados. Estas duas variáveis são:

a) qualitativas b) ambas discretas c) ambas contínuas

d) contínua e discreta, respectivamente

e) discreta e contínua, respectivamente

2) A distribuição abaixo indica o número de acidentes ocorridos em uma empresa com 70 funcionários. (dados fictícios).

Nº de acidentes	0	1	2	3	4	5	6	7
Nº de funcionários	20	10	16	9	6	5	3	1

Determine:

a) o número de funcionários que não sofreram acidente;

b) o número de funcionários que sofreram pelo menos 4 acidentes;

c) o número de funcionários que sofreram $1 < \text{acidentes} \leq 4$;

d) o número de funcionários que sofreram no mínimo 3 e no máximo 5 acidentes;

e) a porcentagem dos funcionários que sofreram no mínimo 5 acidentes;

f) a porcentagem dos funcionários que sofreram entre 2 e 4 acidentes;

g) gráficos de colunas e de barras.

3) Os depósitos bancários da Empresa AKI-SE-TRABALHA, em milhares de Reais, Fev/Mar, 2005:

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

a) Ordenar os dados pelo dispositivo ramo e folhas. (também pelo computador).

b) Construa a distribuição de frequências usando as técnicas de construção.

c) Faça o histograma, o polígono de frequência e a ogiva do item b.

4) Se os salários dos professores do Estado aumentam em 20% em dado período, enquanto o Índice de Preços aumenta em 10%, então, o aumento real de salário, durante o período, foi:

a) de 10% b) maior que 10% c) menor que 10% d) nulo

5) Substituir por uma tabela o trecho do relatório seguinte retirado do IBGE - Estatísticas de Registro Civil 2004. No Brasil, a porcentagem de óbitos violentos para indivíduos do sexo masculino entre 2000 e 2003, nas Regiões; Norte, Nordeste, Sudeste, Sul e Centro Oeste são: 2000 – Norte 17,4%, Nordeste 13,4%, Sudeste 17,3%, Sul 13,6% e Centro-Oeste 19,6%; 2001 – Norte 17,6%, Nordeste 13,5%, Sudeste 17,4%, Sul 14,6% e Centro-Oeste 19,4%; 2002 – Norte 17,5%, Nordeste 13,4%, Sudeste 17,5%, Sul 13,5% e Centro-Oeste 19,5%; 2003 – Norte 15,8%, Nordeste 13,6%, Sudeste 17,0%, Sul 13,3% e Centro-Oeste: 19,7%. *Construir também o gráfico de colunas.*

6) Substituir por uma tabela o trecho do relatório seguinte retirado do IBGE - Estatísticas de Registro Civil 2004. No Brasil, a porcentagem de óbitos violentos para indivíduos do sexo masculino é quase 4 vezes superior à do sexo feminino. Baseado em dados existentes entre 2000 e 2003, a situação no Norte, Nordeste, Sudeste, Sul e Centro Oeste é a seguinte: 2000 – Norte: 17,4% masculino e 5,8% feminino; Nordeste: 13,4% masculino e 3,8% feminino; Sudeste: 17,3% masculino e 4,4% feminino; Sul: 13,6% masculino e 4,4% feminino e Centro-Oeste: 19,6% masculino e 6,5% feminino; 2001 – Norte: 17,6% masculino e 5,9% feminino; Nordeste: 13,5% masculino e 3,8% feminino; Sudeste: 17,4% masculino e 4,3% feminino; Sul: 14,6% masculino e 5,1% feminino e Centro-Oeste: 19,4% masculino e 6,4% feminino; 2002 – Norte: 17,5% masculino e 5,8% feminino; Nordeste: 13,4% masculino e 3,7% feminino; Sudeste: 17,5% masculino e 4,2% feminino; Sul: 13,5% masculino e 5,7% feminino e Centro-Oeste: 19,5% masculino e 6,3% feminino; 2003 – Norte: 15,8% masculino e 4,7% feminino; Nordeste: 13,6% masculino e 3,4% feminino; Sudeste: 17,0% masculino e 4,3% feminino; Sul: 13,3% masculino e 3,6% feminino e Centro-Oeste: 19,7% masculino e 6,0% feminino.

7) Um professor preencheu uma tabela, enviado pelo Departamento de Educação, com os seguintes dados:

Série e Turma	Nº de alunos 30/03	Nº de alunos 30/11	Promovidos sem recuperação	Retidos sem Recuperação	Em recuperação	Recuperados	Não Recuperados	Total Geral	
								Promovidos	Retidos
1º B	49	44	35	03	06	05	01	40	04
1º C	49	42	42	00	00	00	00	42	00
1º E	47	35	27	00	08	03	05	30	05
1º F	47	40	33	06	01	00	01	33	07
Total	192	161	137	09	15	08	07	145	16

Pede-se:

a) a taxa de evasão, por classe;

b) a taxa de evasão total;

c) a taxa de aprovação, por classe;

d) a taxa de aprovação geral;

e) a taxa de recuperação, por classe;

f) a taxa de recuperação geral;

g) a taxa de reprovação na recuperação geral;

h) a taxa de aprovação, sem a recuperação;

i) a taxa de retidos, sem a recuperação.

8) A tabela abaixo apresenta uma distribuição de frequência das áreas de 400 lotes:

Áreas (m^2)	300 |-- 400	|--500	|-- 600	|-- 700	|--800	|--900	|-- 1.000	|-- 1.100	|-- 1.200
Nº de Lotes	14	46	58	76	68	62	48	22	6

Determine:

a) o limite inferior da quinta classe

b) o ponto médio da sétima classe

c) a amplitude do intervalo da sexta classe

d) a frequência da quarta classe

e) a frequência relativa da sexta classe

f) a freq. acumulada da quinta classe

g) o número de lotes cuja área não atinge 700 m^2.

h) o número de lotes igual ou maior a 800 m^2.

i) a porcentagem dos lotes cuja área não atinge 600 m^2.

j) a porcentagem dos lotes cuja área é de 500 m^2, no mínimo, mas inferior a 1.000 m^2.

6 MEDIDAS ESTATÍSTICAS

Além da construção de tabelas e gráficos, a análise exploratória de dados, consiste também de cálculos de medidas estatísticas que resumem as informações obtidas dando uma visão global dos dados. Essas medidas, também conhecidas como medidas descritivas, recebem o nome genérico de estatísticas quando calculada com os dados da amostra, e de parâmetros quando calculadas com dados populacionais.

Dentre as medidas estatísticas as mais utilizadas são as *de tendência central (*ou de *posição) e* as de *dispersão (*ou de *variabilidade).* Destacam-se, ainda, *as separatrizes, as assimetrias e os box plot.*

6.1 MEDIDAS TENDÊNCIA CENTRAL (POSIÇÃO)

As medidas de tendência central são aquelas que produzem um valor em torno do qual os dados observados se distribuem, e que visam sintetizar em um único número o conjunto de dados. As medidas de tendência central são: média aritmética, mediana e moda.

6.1.1 Média

Uma das medidas estatísticas mais utilizadas na representação de uma distribuição de dados é a média aritmética, na sua forma simples, ou ponderada. No primeiro caso divide-se a soma de todos os valores da série pelo número de observações, enquanto no segundo, mais utilizado em distribuições de frequências, os valores são ponderados pelas frequências com que ocorrem e depois dividem-se pelo total das frequências (*este segundo caso será visto em distribuição de frequências*):

Simples: $\overline{X} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$ ou simplesmente $\overline{X} = \frac{\sum x_i}{n}$

Exemplo: Foram levantados os diâmetros de 10 peças (cm) da Empresa AA Ltda. As medidas foram as seguintes: 13,1 – 13,5 – 13,9 – 13,3 – 13,7 – 13,1 – 13,1 – 13,7 – 13,2 – 13,5. Portanto, diâmetro médio é ***13,41 cm.***

A média aritmética possui algumas propriedades desejáveis e não desejáveis e são as seguintes:

i. Unicidade. Para um conjunto de dados existe somente uma média aritmética.
ii. Simplicidade. A média aritmética é fácil de ser interpretada e de ser calculada.
iii. Todos os valores entram para o cálculo da média aritmética, porém, os valores extremos afetam no valor calculado, e em alguns casos pode haver uma grande distorção, tornando, neste caso, a média aritmética indesejável como medida de tendência central.

Como a média é influenciada por valores extremos da distribuição, ela só deve ser utilizada em distribuições simétricas, ou levemente assimétricas, e em distribuições não heterogêneas. Sua aplicação nos dois casos acima é precária e de pouca utilidade prática, pois perde sentido prático e capacidade de representar a distribuição que a originou.

Também nos casos de série em que o fenômeno tem uma evolução não linear, como as séries de valores financeiros no tempo, de acordo com uma capitalização composta, a média mais recomendada seria a *geométrica*. Finalmente, não se recomenda à aplicação da média aritmética nas séries cujos valores representem relações recíprocas, como por exemplo, velocidades, expressas através da relação entre o espaço e o tempo. Neste último caso recomenda-se a utilização da *média harmônica*.

6.1.2 Mediana

A mediana é o valor que ocupa a posição central de um conjunto de valores ordenados, ou seja, medida divide a distribuição de valores em duas partes iguais: 50% acima e 50% abaixo do seu valor. Quando o conjunto possui quantidade par de valores, há dois valores centrais, neste caso, a mediana é o valor médio dos dois valores centrais do conjunto de dados ordenados.

Exemplo: Com os dados do exemplo anterior, calcular a mediana.

13,1 – 13,1 – 13,1 – 13,2 – **13,3 – 13,5** – 13,5 – 13,7 – 13,7 – 13,8

Nesta série tem-se número par de observações logo, têm-se dois valores centrais e são 13,3 e 13,5. Logo, a mediana é ***13,4 cm.***

Suponha, neste mesmo exemplo que se acrescente o valor 14,0 tornando um rol de número ímpar,

13,1 – 13,1 – 13,1 – 13,2 – 13,3 – **13,5** – 13,5 – 13,7 – 13,7 – 13,8 – 14,0

Neste caso, a série possui apenas um valor central logo, a mediana é igual a ***13,5 cm***.

Propriedades da mediana

i. Unicidade. Existe somente uma mediana para um conjunto de dados.
ii. Simplicidade. A mediana é fácil de ser calculada.
iii. A mediana não é tão afetada pelos valores extremos como a média aritmética, por isso, se diz que a mediana é uma medida robusta.

Conceito de resistência de uma medida

Diz-se que uma medida de centralidade ou de dispersão é resistente quando ela é pouco afetada pela presença de observações discrepantes. Entre as medidas de centralidade, a média é bem menos resistente que a mediana. Por outro lado, entre as medidas de dispersão, o desvio padrão é bem menos resistente do que o desvio inter-quartílico.

6.1.3 Moda

Moda de um conjunto de valores é o valor que ocorre com maior frequência, sua aplicação não depende do nível de mensuração da variável, sendo aplicada tanto a fenômenos qualitativos quanto quantitativos. Se todos os valores forem diferentes não há moda, por outro lado, um conjunto pode ter mais do que uma moda: bimodal, trimodal ou multimodal.

Exemplo: Para os dados dos exemplos anteriores a moda é igual a ***13,1 cm***.

A moda pode ser utilizada para descrever dados qualitativos. Por exemplo, suponha que os pacientes vistos em uma clínica de saúde mental durante um determinado ano receberam um dos seguintes diagnósticos: retardo mental, psicose,

neurose e mudança de personalidade. O diagnóstico que ocorre com maior frequência no grupo de pacientes pode ser chamado de *diagnóstico modal.*

6.2 MEDIDAS DE DISPERSÃO

A dispersão de conjunto de dados é a variabilidade que os dados apresentam entre si. Se todos os valores forem iguais, não há dispersão; se os dados não são iguais, existe dispersão entre os dados. A dispersão é pequena quando os valores são próximos uns dos outros. Se os valores são muito diferentes entre si, a dispersão é grande, assim, as medidas de dispersão apresentam o grau de agregação dos dados. Veja como exemplo a Tabela 14.

Tabela 14: Valores das séries A, B e C

Repetição	Série A	Série B	Série C
1	45	41	25
2	45	42	30
3	45	43	35
4	45	44	40
5	45	45	45
6	45	46	50
7	45	47	55
8	45	48	60
9	45	49	65
Média	**45**	**45**	**45**
Mediana	**45**	**45**	**45**

Nota-se que a série “A” não apresenta dispersão, já os valores da série “B” apresentam certa dispersão em torno da média 45, e os valores da série “C” apresentam uma dispersão em torno da média e maior do que a da série B.

As medidas descritivas mais comuns para quantificar a dispersão são: *amplitude, desvio médio, variância, desvio-padrão e coeficiente de variação.*

6.2.1 Amplitude

Uma maneira de medir a variação em um conjunto de valores é calcular a amplitude. A amplitude é a diferença entre o maior e o menor valor de um conjunto de observações.

$$At = n^{o} \text{ maior} - n^{o} \text{ menor}$$

Exemplo: Determinar amplitude total da série: A, B e C.

A utilidade da amplitude total como medida de dispersão é muito limitada, pois depende apenas dos valores extremos. A maior vantagem em usá-la é a simplicidade do seu cálculo.

6.2.2 Desvio Médio

Uma vez que se deseja medir a dispersão ou grau de concentração dos valores em torno da média, nada mais interessante do que analisar o comportamento dos desvios de cada valor em relação à média, isto é:

$$d_i = (x_i - \bar{x})$$

Porém, para qualquer conjunto de dados, a soma de todos os desvios é igual a zero, isto é:

$$\sum d_i = \sum (x_i - \bar{x}) = 0$$

Neste caso, considera-se o módulo de cada desvio $\left|x_i - \bar{x}\right|$, evitando com isso que $\sum d_i = 0$. Dessa forma, o desvio de um conjunto de n valores é dado por:

$$DM = \frac{\sum_{i=1}^{n} \left|x_i - \bar{x}\right|}{n}$$

Exemplo: Determinar desvio médio da série B.

6.2.3 Variância

Embora o desvio médio seja uma medida melhor do que a Amplitude, ainda não é uma medida ideal, pois não discrimina pequenos dos grandes afastamentos em relação à média. Se para eliminar o problema dos sinais, ao invés de considerarmos os valores absolutos elevarmos os afastamentos ao quadrado, estaremos não apenas eliminando o problema dos sinais como também potencializando os afastamentos, enfatizando os grandes desvios em relação às

observações mais próximas da média. Como resultado define a medida de variação, denominada de variância, como:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \qquad \text{ou} \qquad s^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}{n-1}$$

Exemplo: Determinar as variâncias das séries A, B e C.

Esta estatística isolada tem difícil interpretação por apresentar unidade de medida igual ao quadrado da unidade de medida dos dados.

6.2.4 Desvio Padrão

Devido à dificuldade de interpretação da variância, por ter sua unidade de medida ao quadrado, na prática usa-se o desvio padrão que é a raiz quadrada da variância, ou seja:

$$s = \sqrt{s^2}$$

Exemplo: Determinar os desvios-padrão das séries A, B e C.

6.2.5 Erro Padrão

Diferentes amostras retiradas de uma mesma população podem apresentar médias diferentes. A variação existente entre este conjunto de médias é estimada através do erro padrão, que corresponde ao desvio padrão das médias, sendo representado por $s_{\bar{x}}$ e calculado pela fórmula:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

6.2.6 Coeficiente de Variação

Uma pergunta que pode surgir é se um desvio-padrão é grande ou pequeno; questão relevante, por exemplo, na avaliação da precisão de métodos. Um desvio-padrão pode ser considerado grande ou pequeno dependendo da ordem de grandeza da variável. Por exemplo, um desvio-padrão de 10 pode ser insignificante

se a observação típica for 10.000, mas será um valor bastante significativo para um conjunto de dados cuja observação típica é 100.

O coeficiente de variação é uma medida relativa de dispersão, utilizada para comparar, em termos relativos, o grau de concentração em torno da média. É representada por:

$$CV = \frac{s}{\overline{X}}$$

O CV é uma medida adimensional, isto é, sem unidade de medida, podendo ser expressa em termos decimais ou percentuais (multiplicando por 100). Dizemos que uma distribuição é homogênea quando a variabilidade relativa expressa pelo coeficiente de variação, não ultrapassar a 20% . Obviamente a distribuição não deixa de ser homogênea para valores maiores do que 20% mas vai perdendo o grau de homogeneidade na medida em que o coeficiente aumenta.

Exemplo: Determinar o erro padrão e o coeficiente de variação das séries A, B e C.

Esta medida pode ser bastante útil na comparação de duas variáveis ou dois grupos que a princípio não são comparáveis (por exemplo, com ordens de grandeza das variáveis diferentes).

Exemplo: Comparação dos depósitos bancários de duas Empresas (milhares R$).

A Empresa X depositou, em média mensal, 2,0 (milhares R$) e um desvio-padrão de 0,5 (milhares R$). A Empresa Y depositou média mensal, 2,3 (milhares R$) e um desvio-padrão de 0,8 (milhares R$). A Empresa Y apresenta não só uma média mensal mais alta como também maior variabilidade em torno da média. O coeficiente de variação capta esta diferença. Neste caso, o coeficiente de variação é 25% para a Empresa X e 34,8% para a Empresa Y.

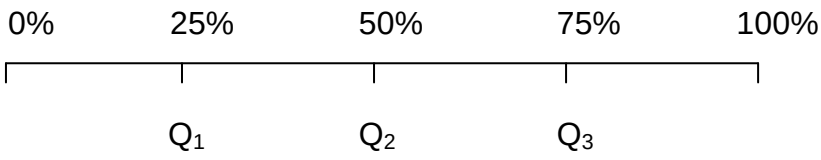
Alguns especialistas consideram:

- Baixa dispersão: CV ≤ 15%
- Média dispersão: 15% < CV < 30%
- Alta dispersão: CV ≥ 30%.

6.3 SEPARATRIZES: QUARTIS, DECIS E PERCENTIS

Os quartis, decis e percentis são muito similares à mediana, uma vez que também subdividem a distribuição de medidas de acordo com a proporção das frequências observadas.

Os quartis dividem um conjunto de dados em quatro partes iguais, isto é, 25% por parte.

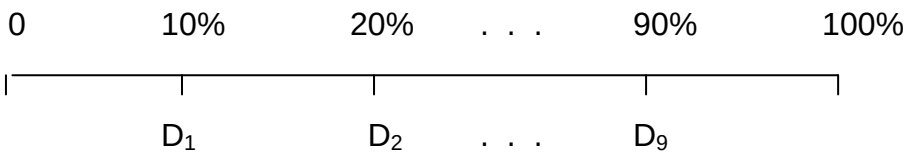


onde: Q_1 = 1^0 quartil, deixa 25% dos elementos.

Q_2 = 2^0 quartil, deixa 50% dos elementos (coincide com a mediana).

Q_3 = 3^0 quartil, deixa 75% dos elementos.

Os decis dividem um conjunto de dados em dez partes iguais, isto é, 10% por parte.



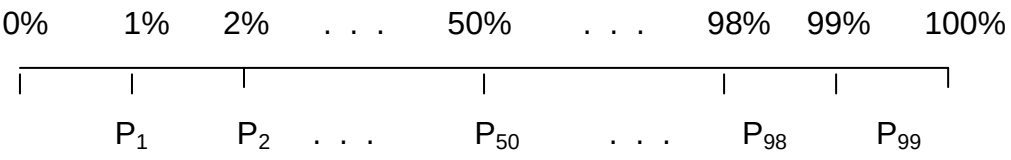
onde: D_1 = 1^0 decil, deixa 10% dos elementos.

D_2 = 2^0 decil, deixa 20% dos elementos.

.....

D_9 = 9^0 decil, deixa 90% dos elementos.

Já, os percentis permitem dividir o conjunto de dados em 100 partes, sendo e 1% em cada parte.



onde: P_1 = 1^0 percentil, deixa 1% dos elementos.

P_2 = 2^0 percentil, deixa 2% dos elementos.

.....

P_{99} = 99^0 percentil, deixa 99% dos elementos.

A mediana é o percentil de ordem 50. Pois, a mediana é um valor que divide o conjunto de dados em duas partes iguais, ou seja, 50% dos dados ficam abaixo e 50% acima.

Os percentis de ordem 25, 50 e 75 são os respectivamente primeiro, segundo e terceiro quartis, porque dividem a distribuição em 1/4, 2/4 = 1/2 e 3/4. Logo o Q_2 é outra notação para a mediana.

Enquanto que os decis $D_1, D_2,,...,D_9$ são os valores que dividem o conjunto em dez partes iguais, que coincidem com os percentis $P_{10}, P_{20},,...,P_{90}$, que também dividem os dados em grupos com 10% em cada um. Portanto, os quartis e os decis estão inseridos nos percentis.

Para determinar o valor correspondente a um certo quartil, decil ou percentil, deve seguir a seguinte sequência:

- Ordenar os dados do menor para o maior.
- Localizar a posição (*L*), dado por:

$$L = \frac{k.n}{100}$$

onde: *k* é o percentual desejado e *n* é o número de valores do conjunto de dados.

Se o valor de *L* for *decimal*, arredonda o seu valor para o maior inteiro mais próximo, e quando o valor de *L* for *inteiro*, deve-se somar o valor correspondente a *L* ao valor de *L*+1 e dividir o resultado por 2.

Considere os depósitos bancários da Empresa AKI-SE- TRABALHA, em milhares de Reais, Fev/Mar, 2005, fica:

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

Por exemplo: O percentil 25 que corresponde ao primeiro quartil, que deixa 25% dos dados abaixo e 75% dos dados acima dele, usa-se:

O percentil de ordem 25 (P_{25}) que deixa 25% dos dados abaixo é:

$$L = \frac{25\ x\ 43}{100} = 10{,}75$$ (11º, é aposição que ocupa no conjunto).

Então, P_{25} = 1,6 (que é igual ao primeiro quartil, isto é Q1 = 62,5).

Isto implica que 25% dos depósitos bancários da empresa são iguais ou abaixo de 1,6 (milhares de reais).

6.4 ASSIMETRIA

Embora as medias de posição e de variação possibilitam descrever estatisticamente um conjunto de dados, é necessário verificar como está se comportando de forma geral essa distribuição, o que é possível através da distribuição de frequência e de histograma. Sendo que as distribuições possam tomar praticamente qualquer forma, a maioria que se encontra na prática é discreta por alguns tipos – padrão.

É de suma importância que a distribuição seja em forma de sino, ou seja, é uma distribuição simétrica, pois metade da esquerda do seu histograma é aproximadamente a imagem-espelho da metade direita.

As distribuições consideradas assimétricas apresentam uma “cauda” em uma das extremidades, quando está à direita, é positivamente assimétrica, e se está à esquerda, é negativamente assimétrica.

As distribuições consideradas assimétricas apresentam uma “cauda” em uma das extremidades, quando está à direita, é positivamente assimétrica, e se está à esquerda, é negativamente assimétrica. Para verificar o tipo e o grau da assimetria da distribuição utiliza-se a medida estatística adimensional denominada de Coeficiente de Assimetria de Pearson, definido como:

$$As = \frac{3(\bar{x} - Md)}{s}$$

Para uma distribuição perfeitamente simétrica, o valor de ***As*** é zero, de modo geral, os valores **As** situam-se entre –3 e 3.

Se, $0{,}15 < |As| < 1$, a assimetria é considerada *moderada*; se $|As| > 1$, é *forte*.

Em uma *distribuição simétrica*, a média ($\bar{x}$), a mediana (Md) e a moda (Mo) são iguais, isto é, $\bar{x}$ = Md = Mo. Em uma *distribuição assimétrica positiva* ou *assimétrica à direita*, a média é maior que a mediana, e esta, por sua vez, maior que a moda ($\bar{x}$ > Md > Mo), ao passo que, em uma *distribuição assimétrica negativa* ou *assimétrica à esquerda*, a média é menor que a mediana, e esta, menor que a moda ($\bar{x}$ < Md < Mo). A Figura 16 apresenta um esquema dessas distribuições:

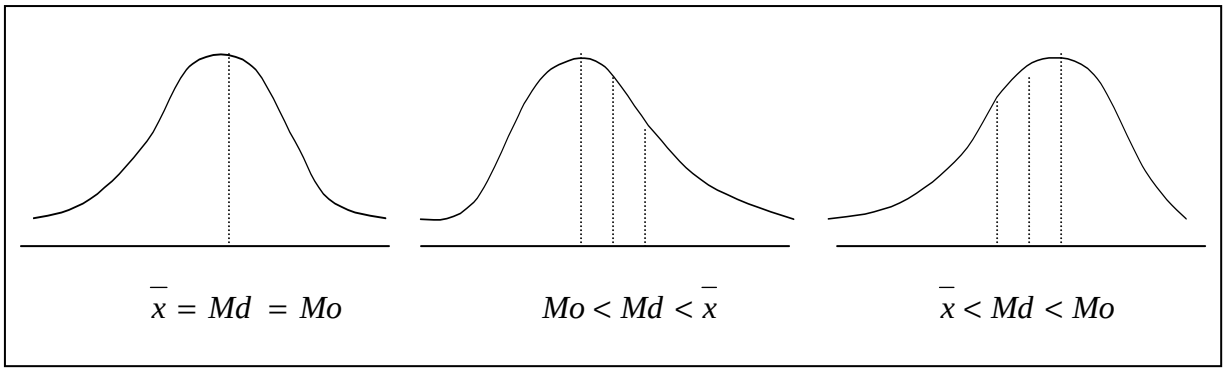


Figura 16: gráficos simétrico e assimétrico à direita e à esquerda

6.5 CURTOSE

Curtose é o grau de achatamento de uma distribuição em relação a uma distribuição padrão, denominada de curva normal.

A curva normal, que é nossa base referencial, recebe o nome de *mesocúrtica*. Já, uma distribuição que apresentar uma curva de frequência mais achatada do que a normal é denominada de *leptocúrtica*, e a que apresentar uma curva de frequência mais aberta, recebe o nome de *platicúrtica*. A Figura 17 apresenta um esquema dessas curvas.

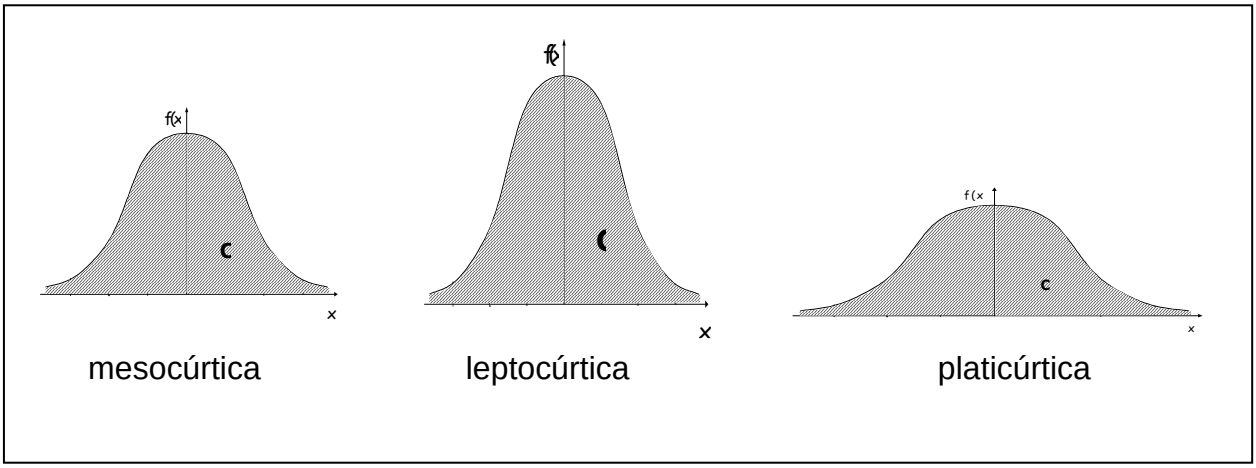


Figura 17: Classificação das curvas em relação a uma distribuição padrão

Para verificar o tipo de curva (da distribuição) e o grau de curtose utiliza-se a medida estatística adimensional denominada de Coeficiente de Curtose definido como:

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

Para uma curva relativamente à normal, tem-se que C = 0,263. Isto é:

Se C = 0,263 ➔ curva mesocúrtica

C < 0,263 ➔ curva leptocúrtica

C > 0,263 ➔ curva platicúrtica

6.6 BOX PLOT

O box plot introduzido pelo estatístico americano John Tukey em 1977 é a forma de representar graficamente os dados da distribuição de uma variável quantitativa em função de seus parâmetros. Os cinco itens ou valores: o menor valor (x_1), os quartis (Q_1, Q_2 e Q_3) e o maior valor (x_n), são importantes para se ter uma idéia da posição, dispersão e assimetria da distribuição dos dados. Na sua construção são considerados os quartis e os limites da distribuição, permitindo uma visualização do posicionamento da distribuição na escala da variável. Para melhor compreensão deste box plot, a Figura 18 apresenta um esquema sintetizado:

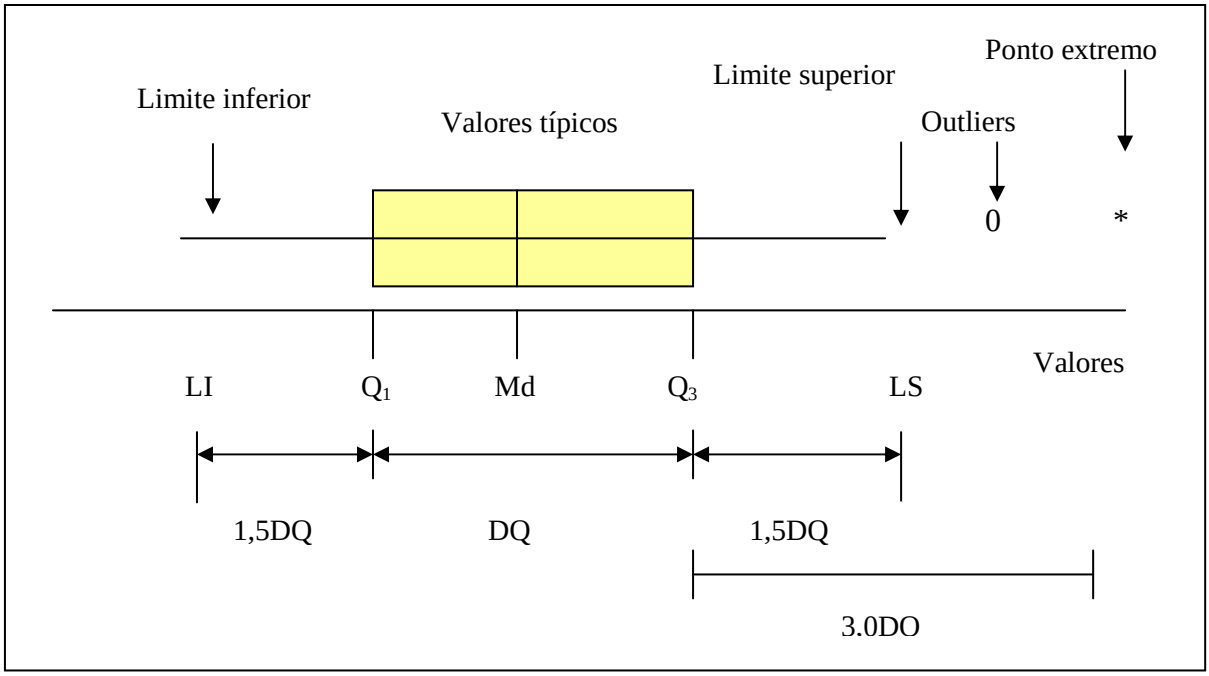


Figura 18: Esquema para construção do b*ox plot*

A escala de medida da variável encontra-se na linha horizontal do quadro onde está inserida a figura.

Na caixa retangular da figura são fornecidos os quartis Q_1, na parte esquerda, e Q_3 na parte direita da caixa. Entre eles encontra-se a Mediana da distribuição. Observe que 50% da distribuição têm valores dentro da caixa.

As linhas horizontais que saem da caixa terminam nos limites inferior (LI) e superior (LS) da distribuição. Entre esses limites encontram-se os valores considerados como típicos da distribuição. Esses limites são determinados em função da distância entre os dois quartis (Q_3 e Q_1), isto é, do desvio inter-quartílico: $DQ = Q_3 - Q_1$.

Observações com afastamento superior a 1,5 desvio inter-quartílico, para cima ou para baixo, são consideradas atípicas, ou possíveis *outliers.* Os pontos que estão mais de 1,5 *DQ* e menos que 3,0 *DQ*, são chamados de outliers*,* aparecendo (o).

Valores com afastamento superior a 3,0 DQ, para cima ou para baixo são considerados como *pontos extremos,* aparecendo na figura com (*). Quanto maior for o valor do desvio inter-quartílico, maior a variabilidade da distribuição.

Obs. Muitos livros e softwares apenas comentam sobre os pontos atípicos chamando-os de **outliers** (pontos discrepantes).

O b*ox plot* também fornece informações importantes sobre o comportamento do conjunto de dados, como simetria e variabilidade. Se a amplitude for muito maior que à distância interquartílica e a mediana estiver mais próxima do 1º quartil do que do 3º quartil, há forte indicação de assimetria positiva e de grande dispersão das observações.

Exemplo: O objetivo da administração é lucrar o máximo possível com o capital investido em sua empresa. Uma medida de bom desempenho é o retorno sobre os investimentos. A seguir são apresentados os mais recentes retornos em milhares (R$).

2.210	2.255	2.350	2.380	2.380	2.390
2.420	2.440	2.450	2.550	2.630	2.825

A mediana é 2.405 e os quartis Q_1 = 2.365 e Q_3 = 2.500. A resenha dos dados mostra um menor valor 2.210 e um maior valor de 2.825. Assim, a regra de cinco itens (números) para os dados de pesos dos recém nascidos é 2.210; 2.365; 2.405; 2.500; 2.825.

Além desses valores, têm-se os limites, inferior que é dado por LI = Q_1–1,5DQ e superior LS = Q_3 + 1,5DQ. No caso, LI = 2.162,5 e LS = 2.702,5. Os dados fora destes limites são considerados *pontos fora da curva*. Neste caso, “o” = 2.825 é um outliers. A Figura 19 apresenta um esquema do box plot com esses resultados:

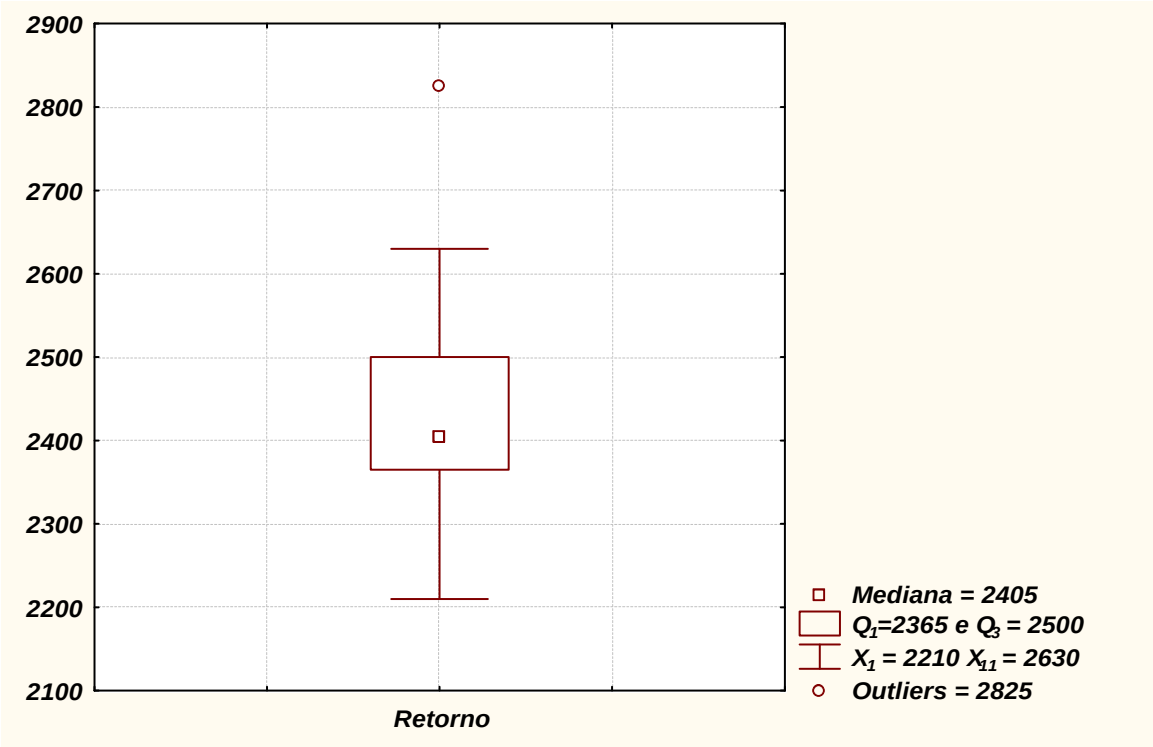


Figura 19: Resultados do desempenho de retorno de investimento da empresa

Observações atípicas *(outlier)*

É muito comum aparecerem entre os dados coletados, observações atípicas (*outliers*), isto é, valores muito grande ou muito pequeno em relação aos demais. Um conjunto de dados pode apresentar apenas um ou vários *outliers*.

Observações atípicas alteram enormemente as médias e variabilidade dos grupos a que pertencem e podem até mesmo distorcer as conclusões obtidas através de uma análise estatística padrão. Portanto, é de fundamental importância detectar e dar um tratamento adequado a elas. É sempre boa a prática fazer-se uma

inspeção dos dados no início da análise estatística. Técnicas descritivas de dados têm um papel importante nesta fase.

Causas do aparecimento de *outliers*

Dentre as possíveis causas do aparecimento de *outliers*, pode citar as seguintes:

- Leitura, anotação ou transição incorreta dos dados.
- Erro na execução do experimento ou na tomada da medida.
- Mudanças não controláveis nas condições experimentais ou dos pacientes.

Como detectar *outliers*

As questões básicas são quais observações devem ser consideradas como *outliers* e como detectá-los. Existem procedimentos para responder a essas perguntas.

Os *outliers* podem ser detectados simplesmente por uma verificação lógica dos dados, através de gráficos específicos ou ainda através de teste apropriados. Uma forma gráfica usual é o b*ox plot*. As plotagens de retângulos são outras maneiras de identificar os pontos fora da curva. Mas eles não necessariamente identificam os mesmos valores que aqueles com uma contagem-z menor que -3 ou maior que +3. No entanto, o objetivo de ambas as abordagens é simplesmente identificar os valores de dados extremos que devem ser revisados para assegurar a validade dos dados. Pontos fora da curva identificados pelos dois métodos devem ser revisados.

6.7 MEDIDAS DE POSIÇÃO E DISPERSÃO DE UMA DISTRIBUIÇÃO DE FREQUÊNCIA

Quando existe uma grande quantidade de dados, estes podem ser agrupados. A finalidade em agrupar os dados é para facilitar os cálculos.

Exemplo: Um novo medicamento para cicatrização está sendo testado e um experimento é feito para estudar o tempo (em dias) de completo fechamento em cortes provenientes de cirurgia. Uma amostra em trinta cobaias forneceu os valores: 15, 17, 16, 15, 17, 14, 17, 16, 16, 17, 15, 18, 14, 17, 15, 14, 15, 16, 17, 18, 18, 17, 15, 16, 14, 18, 18, 16, 15 e 14.

a) Organize uma tabela de frequência.
b) Obter as frequências relativas de cada classe.
c) Calcular a média.
d) Que porcentagem das observações está abaixo de 16 dias?
e) Classifique como *rápida* as cicatrizações iguais ou inferior a 15 dias e como *lenta* as demais. Quais as porcentagens para cada classificação.

Solução: **a** e **b**

Cicatrização	14	15	16	17	18	total
Frequência	5	7	6	7	5	30
Frequência relativa	0,167	0,233	0,200	0,233	0,167	1,000
$x_i . f_i$	70	105	96	119	90	480

Média $$\bar{x} = \frac{\sum x_i \cdot f_i}{n} = \frac{480}{30} = 16$$

A determinação das medidas de posição e de dispersão para uma *variável quantitativa contínua*, através de sua distribuição de frequências, exige aproximações, já que perde a informação dos valores observados. *Por exemplo*, com as quantidades de depósitos bancários (milhares R$), a distribuição de frequência está representada na Tabela 15.

Tabela 15: Nível de ruído, em decibéis, de tráfego em certo cruzamento

Nível de ruído (em db)	**Quantidade** (f_i)	**Ponto médio** ($\bar{x}_i$)	**Freq. Acum.** (F_{ac})	($x_i . f_i$)	($x_i^2 . f_i$)
58,0 |-- 60,0	5	59	5	295	17.405
60,0 |-- 62,0	5	61	10	305	18.605
62,0 |-- 64,0	6	63	16	378	23.814
64,0 |-- 66,0	9	65	25	585	38.025
66,0 |-- 68,0	15	67	40	1.005	67.335
68,0 |-- 70,0	5	69	45	345	23.805
70,0 |-- 72,0	5	71	50	355	25.205
Total	**50**			**3.268**	**214.194**

Como foi dito, no agrupamento dos dados acarreta alguma perda de informação. Cada elemento perde sua identidade, por isso, sabem apenas quantos elementos há em cada classe. Uma aproximação razoável é supor que todos os valores dentro de cada classe tenham seus valores iguais ao ponto médio desta classe.

6.7.1 Média

Para o cálculo da média, em geral, obtém-se uma boa aproximação atribuindo a cada elemento que se enquadra em uma classe o valor médio correspondente. Esse processo em geral é satisfatório, pois os erros introduzidos nos cálculos tendem a compensar-se.

A fórmula para a média de uma distribuição de frequências, onde x_1, x_2, ..., x_n são os valores médios das classes, ponderados pelas frequências correspondentes f_1, f_2, ..., f_n é dada por:

$$\bar{x} = \frac{\sum_{i=1}^{n} \bar{x}_i . f_i}{n}, \quad \text{assim} \qquad \bar{x} = \frac{3.268}{50} = 65,36$$

6.7.2 Mediana

A mediana divide um conjunto de dados ordenados em duas partes iguais. A expressão para determinar a mediana de uma distribuição de frequências é dada por:

$$Md = l_i + \frac{\frac{n}{2} - F_{ac-1}}{f_{Md}}\ a_C, \quad \text{assim} \quad Md = 64 + \frac{25-16}{9}2 = 66,0$$

onde: l_i = limite inferior da classe da mediana; n = número de elementos;

a_C = amplitude da classe;

F_{ac-1} = frequência acumulada anterior à classe da Md;

f_{Md} = frequência simples da classe da Md;

Para isso tem-se que:

1^0) Calcular a posição, isto é, a ordem n/2.

2^0) Identificar a classe que contém a mediana, pela frequência acumulada.

6.7.3 Moda

A moda de um conjunto de n números é o valor que ocorre com maior frequência. A expressão para determinar a moda de uma distribuição de frequências é dada por:

$$Mo = l_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}\ a_C, \quad \text{assim} \quad Mo = 66 + \frac{6}{6+10}2 = 66,75$$

Para isso tem que identificar a classe modal (de maior frequência)

l_i = limite inferior da classe modal; a_C = amplitude da classe.

Δ_1 = diferença entre a frequência da classe modal e a anterior;

Δ_2 = diferença entre a frequência da classe modal e a posterior;

Obs. *Pelos cálculos, nota-se que a curva dos dados da tabela é assimétrica à direita, já que a média > mediana > moda*.

6.7.4 Separatrizes: Quartis, Decis e Percentis

a1) Quartis

Os quartis dividem um conjunto de dados em quatro partes iguais. A fórmula para o cálculo dos quartis de uma distribuição de frequência é dada por:

$$Q_i = l_i + \frac{\frac{i\ .\ n}{4} - F_{ac-1}}{fQ_i}\ a_C$$

1^0) Calcula-se $\frac{i\,.\,n}{4}$, onde i = 1, 2 e 3.

2^0) Identifica-se a classe Q_i pela F_{ac}.

a2) Decis

Os decis dividem um conjunto de dados em dez partes iguais. A fórmula para o cálculo dos decis de uma distribuição de frequência é dada por:

$$D_i = l_i + \frac{\frac{i\,.\,n}{10} - F_{ac-1}}{fD_i}\, a_C$$

1^0) Calcula-se $\frac{i\,.\,n}{10}$, onde i = 1, 2, . . . , 9.

2^0) Identifica-se a classe D_i pela F_{ac}.

a3) Percentis

Os percentis dividem um conjunto de dados em cem partes iguais. A fórmula para o cálculo dos percentis de uma distribuição de frequência é dada por:

$$P_i = l_i + \frac{\frac{i\,.\,n}{100} - F_{ac-1}}{fP_i}\, a_C$$

1^0) Calcula-se $\frac{i\,.\,n}{100}$, onde i = 1, 2, . . . , 99.

2^0) Identifica-se a classe P_i pela F_{ac}.

Exemplo: Calcular o percentil de ordem 50 $\quad p_{50} = Md = 64 + \frac{25-16}{9} 2 = 66{,}0$

Como já foi dito, os quartis, decis e percentis são muito similares à mediana, uma vez que também subdividem a distribuição de medidas de acordo com a proporção das frequências observadas.

A mediana é o percentil de ordem 50, já que a mediana é um valor que divide o conjunto de dados em duas partes iguais, ou seja, 50% dos dados ficam abaixo e 50% acima.

Os percentis de ordem 25, 50 e 75 são chamados, respectivamente primeiro, segundo e terceiro quartis porque dividem a distribuição em 1/4, 2/4 e 3/4. São

representados por Q_1, Q_2 e Q_3 e, evidentemente, Q_2 é outra notação para a mediana. Enquanto que os decis D_1, D_2,,...,D_9 são os valores que dividem o conjunto em dez partes iguais, que coincidem com os percentis P_{10}, P_{20},,...,P_{90} , que também dividem os dados em grupos com 10% em cada um. Assim, a fórmula do *percentil* sintetiza as expressões da mediana, dos quartis e dos decis.

6.7.5 Cálculo das Separatrizes Utilizando Proporções

Calcular a mediana utilizando proporções com os dados da Tabela 15. Neste caso constrói-se o histograma com as frequências relativas (Figura 20).

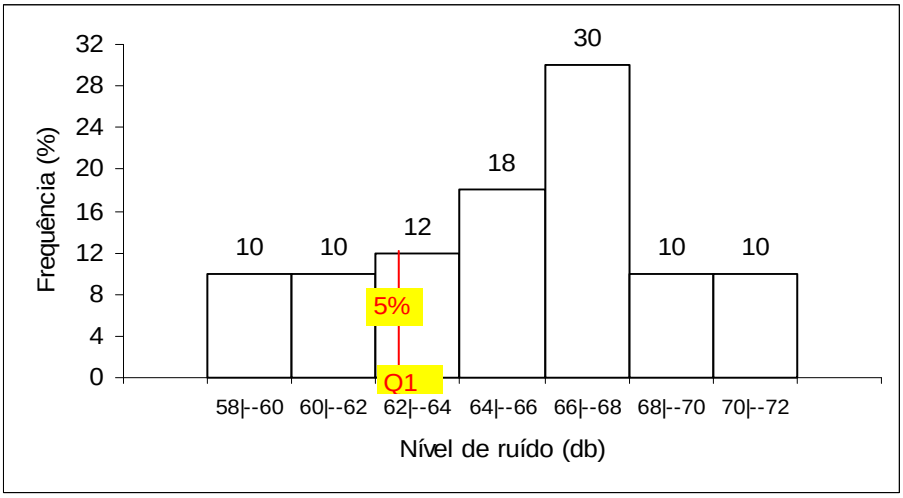


Figura 20: O nível de ruído de certo cruzamento

$$\frac{Q_1 - 62}{5} = \frac{64 - 62}{12} \quad ==> \quad Q = 62{,}83$$

Exemplo: A Tabela 16 apresenta as frequências relativas de ocorrências de faixas de altura (em cm) para uma amostra de 100 crianças de 12 anos de idade.

Tabela 16: Altura de 100 crianças

Faixas	Frequência relativa
100 | -- 110	0,10
110 | -- 120	0,25
120 | -- 130	0,30
130 | -- 140	0,25
140 | -- 150	0,10

a) Construa o histograma

b) Calcule a mediana

c) Desejando-se separar as 15 crianças mais altas, qual seria o ponto de corte?

6.7.6 Desvio Médio

O desvio médio para dados agrupados, isto é, de uma distribuição de frequências é calculado da seguinte forma:

$$DM = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right| f_i}{n} \qquad \text{e} \qquad \bar{x} = \frac{\sum x_i \ f_i}{n}$$

onde: x_i são os pontos médios das classes e os f_i as respectivas frequências.

6.7.7 Variância

A expressão para o cálculo da variância amostral de uma distribuição de frequências é:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 \ f_i \ - \ \frac{(\sum_{i=1}^{n} x_i \ f_i)^2}{n}}{n-1}$$

Obter a variância referenta a tabela 20.

$$s^2 = \frac{214194 - \frac{(3268)^2}{50}}{50-1} = 12{,}19$$

6.7.8 Desvio Padrão

O desvio padrão é obtido extraindo a raiz quadrada da variância, isto é:

$$s = \sqrt{s} \quad ==> \quad s = \sqrt{12{,}94} = 3{,}49$$

6.7.9 Erro Padrão

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{3{,}49}{\sqrt{50}} = 0{,}49$$

6.8 LISTA 2 - EXERCÍCIOS

1) Considere os seguintes dados amostrais (conjunto de peças, em gramas):

100 – 105 – 110 – 102 – 103 – 107 – 105 – 90 – 80

a) Pede-se: a média, a mediana, a moda, o desvio médio, a variância, o desvio padrão, o erro padrão, e o coeficiente de variação.

b) Os dados possuem pequena dispersão? Por quê?

c) Somar 100 de cada observação para obter uma amostra com valores transformados e calcule a média, a variância. (Compare essa variância com os dados originais).

2) Os coeficientes de liquidez obtidos da análise de balanço em 60 indústrias são apresentados em forma ordenada abaixo.

4,44	4,47	4,50	4,54	4,61	4,64	4,67	4,69	4,70	4,75
4,76	4,79	4,81	4,84	4,86	4,87	4,90	4,92	4,95	4,97
4,97	5,00	5,01	5,03	5,05	5,08	5,08	5,09	5,11	5,11
5,12	5,14	5,15	5,17	5,18	5,20	5,22	5,23	5,25	5,26
5,28	5,30	5,32	5,33	5,34	5,36	5,39	5,40	5,41	5,43
5,45	5,47	5,50	5,55	5,59	5,63	5,68	5,72	5,80	5,85

Pede-se:

a) a média; b) a mediana; c) o primeiro quartil;

d) o quinto decil; e) o vigésimo quinto percentil;

f) o desvio-padrão (usar calculadora); h) o coeficiente de variação;

i) é uma distribuição simétrica ou assimétrica (positiva ou negativa)? Justifique.

j) o coeficiente de curtose. Explicar o tipo da curva.

l) explicar os resultados dos quartis, decis e percentis;

3) Em certo ano, além de outros remédios uma farmácia vendeu quatro tipos relevantes. Vendeu 450 remédios da marca X por R$ 120,00 cada um, 350 da marca Y por R$ 130,00 cada um, 220 da marca Z por R$ 145,00 cada um e 180 da marca W por R$ 95,00 cada um de seus. Qual o valor médio desses quatro tipos de remédios vendidos?

4) Em um exame de colesterol, o grau médio de um grupo “A” de 150 pessoas foi de 214 *mg/dl* e um desvio-padrão de 22 *mg/dl*. Em um outro grupo “B”, entretanto, grau médio de 150 pessoas foi de 201 *mg/dl* e um desvio-padrão de 21 *mg/dl.* Em que grupo foi maior a dispersão?

5) Cronometrando o tempo para várias provas de uma gincana automobilística, encontrou-se:

Equipe 1:

8 provas

Tempo médio: 15 segundos

Variância 22 segundos^2

Equipe 2:

Tempo:	10	15	20	25
Nº de provas:	3	2	3	2

Pede-se: a) Qual o coeficiente de variação relativo à equipe 1?

b) Qual o tempo médio e o desvio padrão da equipe 2?

c) Qual a equipe que apresentou resultados mais disperso? Por quê?

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

Nível	Frequência
2,35 |-- 2,55	1
2,55 |-- 2,75	3
2,75 |-- 2,95	2
2,95 |-- 3,15	4
3,15 |-- 3,35	5
3,35 |-- 3,55	6

a) Determine os quartis: 1º., 2º. e 3º. pela fórmula de dados agrupados.

b) Construa o histograma

c) Determine os quartis: 1º., 2º. e 3º. utilizando proporções

d) Qual a porcentagem de valores que estão acima do nível 3?

7) As vendas anuais, em milhões de dólares, para 21 empresas farmacêuticas são apresentadas a seguir:

8.408	1.374	1.872	8.879	2.459	11.413
608	14.138	6.452	1.850	2.818	1.356
10.498	7.478	4.019	4.341	739	2.127
3.653	5.794	8.305			

a) Obter os cinco itens (números) e os limites inferior e superior.

b) Parece haver pontos fora da curva? Qual(is)?

c) As vendas Johnson & Johnson são as maiores na lista, com US$ 14.138 milhões. Suponha que um erro de lançamento tenha sido cometido e que as vendas tenham sido registradas como US$ 41.138 milhões. Neste caso, este valor é um ponto solto (extremo)? Por quê?

7 TRANSFORMAÇÕES DE VARIÁVEIS

Antes de qualquer análise é fundamental que se proceda a um exame dos dados relativos a uma variável, seja ela qualitativa ou quantitativa. Este procedimento é importante como um primeiro contato do analista com a distribuição, além de servir, também, para avaliar a existência de possíveis valores atípicos na distribuição. Se a variável for qualitativa, a concentração de respostas em torno de umas poucas categorias, a existência de células esparsas, com baixa frequência, ou até mesmo o aparecimento de respostas não esperadas, pode indicar algum problema no levantamento dos dados (questão mal formulada ou resposta inválida). No caso da variável ser quantitativa, valores muito afastados da distribuição, ou até mesmo distribuições com assimetria acentuada pode indicar a existência de outliers ou a necessidade de se proceder a uma transformação na escala da variável.

A escolha e a mudança de escalas são artifícios úteis para melhor compreensão de fenômenos. Considere as notas de uma turma de dez alunos em três exames, conforme a Tabela 17:

Tabela 17: Notas de uma turma de 10 alunos em três exames

EXAME	**ALUNOS**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Português	36	35	45	38	40	42	44	46	34	40
Matemática	22	23	17	20	21	19	21	17	22	18
Ciências	10	11	8	9	10	10	11	9	12	10

Fonte: Dados hipotéticos

Sendo a média e a dispersão de cada exame:

Português ➔ média $\mu = 40$ e desvio $\sigma = 4$

Matemática ➔ média $\mu = 20$ e desvio $\sigma = 2$

Ciências ➔ média $\mu = 10$ e desvio $\sigma = 1$

Em primeiro lugar, note que as notas de cada exame estão expressas em escalas diferentes. Como consequência, nada se pode comparar o desempenho dos alunos nos três exames. Tampouco pode comparar os desempenhos entre os alunos, o que impede um ordenamento baseado em suas performances.

7.1 MUDANÇA DE ORIGEM

Por uma questão de conveniência, pode-se proceder a uma transformação que separe os escores observados de uma distribuição a partir do seu valor médio. Nesses casos, valores acima da média serão positivos, enquanto aqueles que estiverem abaixo dela serão negativos. A média, como valor central de uma distribuição, passa a ser, desse modo, a origem da nova escala dos escores. No exemplo dos escores nos três exames, essa transformação permite a avaliação dos alunos com respeito ao desempenho individual tendo a média como base.

Na prática, essa transformação está simplesmente movendo toda a distribuição para a direita ou esquerda, dependendo do sinal da média, sem alterar a unidade das medidas, expressa pela mesma unidade de medida da variável. A mudança da origem, de zero para a média é expressa por: $X_i - \mu$, para i = 1, 2, ... , n.

O valor nulo na nova escala verifica-se para os valores da distribuição, na escala primitiva, iguais à média. A Tabela 18 apresenta os escores dos alunos (do exemplo acima) na nova escala. Os valores nessa tabela são expressos em afastamentos, em pontos, da média.

Tabela 18: Valores expressos em relação aos afastamentos, em pontos, da média

EXAME	**ALUNOS**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Português	-4	-5	5	-2	0	2	4	6	-6	0
Matemática	2	3	-3	0	1	-1	1	-3	2	-2
Ciências	0	1	-2	-1	0	0	1	-1	2	0

A tabela 18 permite separar, para cada exame, os alunos que tiveram desempenho superior ou inferior às respectivas médias. Como afastamentos em torno da média, a soma dos novos escores é igual a zero. As unidades não foram alteradas, o que não permite, ainda, comparar os desempenhos entre os exames. Por exemplo, não pode avaliar se o aluno 3 teve um desempenho mais fraco em Matemática ou Ciências. Para isso será necessário colocar as três distribuições numa unidade comum.

7.2 MUDANÇA DA UNIDADE

A transformação acima desloca as distribuições ao longo do eixo das escalas das variáveis, centrando as distribuições num ponto comum (zero). Não obstante, essa transformação preserva as suas unidades originais. Ao dividir os escores de cada distribuição pelos respectivos desvios padrões, estão unificando também as novas unidades das variáveis. A nova unidade de cada distribuição fica, então, expressa em termos das unidades de desvios de cada distribuição. Desse modo, um aluno que tenha obtido 44 pontos num exame cuja média tenha sido de 40 pontos e desvio padrão de 4 pontos, passa a ter 1 unidade de desvio (não mais pontos) acima da média na nova escala. A nova transformação pode ser expressa através de:

$$Z_i = \frac{X_i - \mu}{\sigma}.$$

Tanto a mudança da origem como a da unidade pode ser feita separadamente, mas quando feitas simultaneamente unifica as escalas, que terão média 0 e desvio padrão 1. Por isso, essa transformação é denominada *padronização* dos escores.

Os escores padronizados para as distribuições das notas dos alunos nos três exames do exemplo acima são apresentados na Tabela 19.

Tabela 19: Escores padronizados das notas dos alunos nos três exames

EXAME	**ALUNOS**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Português	-1	-1,25	1,25	-0,5	0	0,5	1	1,5	-1,5	0
Matemática	1	1,5	-1,5	0	0,5	-0,5	0,5	-1,5	1	-1
Ciências	0	1	-2	-1	0	0	1	-1	2	0

Agora sim, pode analisar os escores dos alunos em termos comparativos. Note, por exemplo, que embora o aluno 3 tivesse ficado com 3 pontos abaixo da média em Matemática e 2 pontos abaixo da média em Ciências, o seu desempenho pior foi no exame de Ciências, em que ficou 2 unidades de desvio abaixo da média, tendo sido o aluno de pior performance nessa disciplina, dentre os dez alunos que se submeteram ao exame. Isto significa que análises comparativas devem considerar parâmetros relativos e não absolutos.

8 ANÁLISE BIDIMENSIONAL

8.1 INTRODUÇÃO

Até agora foi visto como organizar e resumir informações pertinentes a uma única variável de um conjunto de dados, mas freqüentemente está interessado em analisar o comportamento conjunto de duas ou mais variáveis aleatórias. Os dados aparecem na forma de uma matriz, usualmente com as colunas indicando as variáveis e as linhas os indivíduos (ou elementos). A Tabela 3 (dados hipotéticos da Companhia MB) apresenta uma matriz com 6 variáveis e 36 indivíduos.

O objetivo principal das análises nessa situação é explorar relações (similaridades) entre as colunas, ou algumas vezes entre as linhas. A distribuição conjunta das frequências será um instrumento poderoso para compreensão do comportamento dos dados.

Inicialmente deter-se-á no caso de duas variáveis ou dois conjuntos de dados e, na sequência, no caso de três variáveis.

Em algumas situações, pode ter dois ou mais conjuntos de dados provenientes da observação da mesma variável. Por exemplo, pode-se estar interessado em comparar os salários dos casados e dos solteiros.

Na Tabela 3 têm-se sete variáveis: *estado civil, grau de instrução, número de filhos, salário, idade e procedência.*

Quando considera duas variáveis ou dois conjuntos de dados, pode ter três situações:

- *as duas variáveis são qualitativas;*
- *as duas variáveis são quantitativas; e*
- *uma variável é qualitativa e a outra é qualitativa.*

As técnicas de análise de dados nas três situações são diferentes. Quando as variáveis são qualitativas, os dados são resumidos em *tabelas de dupla entrada (ou de contingência)*, onde aparecerão as frequências absolutas ou contagens de indivíduos que pertencem simultaneamente a categorias de uma e outra variável; quando as duas variáveis são quantitativas, as observações são provenientes de mensurações e quando se tem uma variável qualitativa e outra quantitativa, em geral analisa-se o que acontece com a variável quantitativa quando os dados são categorizados de acordo com os diversos atributos da variável qualitativa.

8.2 VARIÁVEIS QUALITATIVAS

Suponha que se queira analisar o comportamento conjunto das variáveis: grau de instrução e região de procedência, cujos dados estão contidos na Tabela 3. A distribuição de frequências é representada por uma tabela de dupla entrada como mostra a Tabela 20.

Tabela 20: Distribuição conjunta das frequências das variáveis: grau de instrução e região de procedência

Região de Procedência	**Grau de instrução**			**Total**
	Ensino Fundamental	Ensino Médio	Superior	
Capital	4	5	2	11
Interior	3	7	2	12
Outra	5	6	2	13
Total	12	18	6	36

Fonte: Tabela 3

Cada elemento do corpo da tabela dá a frequência observada das realizações simultâneas das variáveis: *grau de instrução* e *região de procedência*. Dessa forma, nota-se quatro indivíduos da capital com ensino fundamental, sete do interior com ensino médio, etc.

A linha dos totais fornece a distribuição da variável grau de instrução, ao passo que a coluna dos totais fornece a distribuição da variável região de procedência. As distribuições assim obtidas são chamadas tecnicamente de *distribuições marginais*.

Em vez de se trabalhar com frequências absolutas, constrói-se tabelas com frequência relativas. Porém, existem três possibilidades de se expressar as frequências relativas de cada casela (célula).

- em relação ao total geral;
- em relação ao total de cada linha; e
- em relação ao total de cada coluna.

De acordo com o objetivo do problema em estudo, uma delas será a mais conveniente.

A Tabela 21 apresenta a distribuição conjunta das frequências relativas (proporções) com relação ao total geral. Pode-se, então, afirmar que 11,1% dos empregados vêm da capital e têm ensino fundamental. Os totais nas margens fornecem as distribuições unidimensionais de cada uma das variáveis. Por exemplo, 30,6% dos indivíduos vêm da capital, 33,3% do interior e 36,1% de outras regiões.

Tabela 21: Distribuição conjunta das frequências relativas (em porcentagem) em relação ao total geral das variáveis: grau de instrução e região de procedência

Região de Procedência	**Grau de instrução**			**Total**
	Ensino Fundamental	Ensino Médio	Superior	
Capital	11,1%	13,9%	5,6%	30,6%
Interior	8,3%	19,4%	5,6%	33,3%
Outra	13,9%	16,7%	5,6%	36,1%
Total	33,3%	50,0%	16,7%	100,0%

Fonte: Tabela 3

A Tabela 22 a seguir apresenta a distribuição conjunta das frequências relativas com relação ao total das colunas. Pode-se dizer que, entre os empregados com instrução com ensino fundamental (33,3%), médio (27,8%) e superior (33,3%) vêm da capital.

De modo análogo, pode-se construir a distribuição das frequências relativas em relação ao total das linhas.

Tabela 22: Distribuição conjunta das frequências relativas (em porcentagem) em relação aos totais de cada coluna das variáveis: grau de instrução e região de procedência

Região de Procedência	**Grau de instrução**			**Total**
	Ensino Fundamental	Ensino Médio	Superior	
Capital	33,3%	27,8%	33,3%	30,6%
Interior	25,0%	38,9%	33,3%	33,3%
Outra	41,7%	33,3%	33,3%	36,1%
Total	100%	100%	100%	100,0%

Fonte: Tabela 3

A comparação entre as duas variáveis também pode ser feita utilizando-se representações gráficas. A Figura 21 mostra a distribuição da região de procedência por grau de instrução de acordo com os dados da Tabela 22.

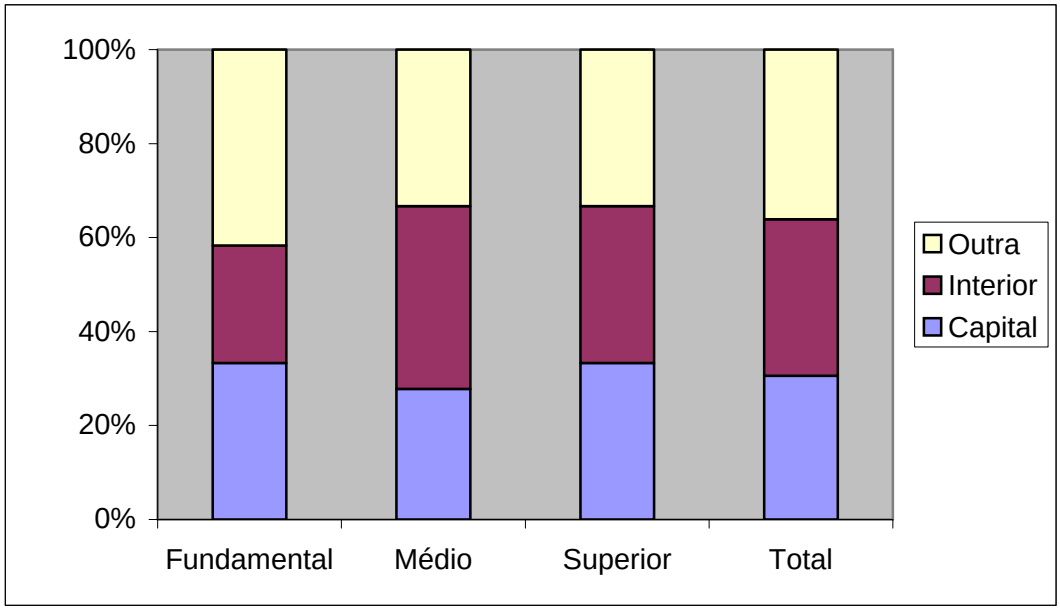


Figura 21: Região de procedência versus grau de instrução

8.3 Associação entre Variáveis Qualitativas

Um dos principais objetivos de se construir uma distribuição conjunta de duas variáveis qualitativas é descrever a associação entre elas, isto é, quando se quer

conhecer o grau de dependência entre elas, de modo que se possa prever o resultado de uma delas quando se conhece a realização da outra. Por exemplo, pode-se estimar a renda média de uma família moradora na cidade de São Paulo, conhecendo a classe social a que ela pertence, pois sabe que existe uma dependência entre as variáveis: renda familiar e classe social.

Para identificar se existe uma associação entre duas variáveis: sexo e carreira escolhida por 200 alunos da distribuição conjunta apresentada na Tabela 23, deve construir as proporções (porcentagens) segundo as linhas ou as colunas para poder fazer comparações.

Tabela 23: Distribuição conjunta de 200 alunos de acordo com sexo e com o curso escolhido

Curso Escolhido	**Sexo**		**Total**
	Masculino	Feminino	
Economia	85	35	120
Administração	55	25	80
Total	140	60	200

Fonte: Dados hipotéticos

A Tabela 24 apresenta as porcentagens, isto é, as frequências relativas referentes ao sexo por curso escolhido, que são obtidas fixando-se os totais das colunas em 100%.

Com os dados da tabela nota-se que, *independentemente do sexo*, 60% das pessoas preferem Economia e 40% Administração (observe na coluna total). Não tendo dependência entre as varáveis, espera essas mesmas porcentagens para cada sexo. Observando a tabela, vê que as porcentagens do sexo masculino (61% e 39%) e do sexo feminino (58% e 42%) são próximas das marginais. Esses resultados parecem indicar que não existe dependência entre as duas variáveis, para o conjunto de alunos considerados. Conclui-se, então, que as variáveis: sexo e escolha do curso não estão *associadas*.

Tabela 24: Distribuição conjunta das porcentagens dos 200 alunos de acordo com sexo e com o curso escolhido

Curso Escolhido	**Sexo**		**Total**
	Masculino	Feminino	
Economia	61%	58%	60%
Administração	39%	42%	40%
Total	100%	100%	100%

Fonte: Tabela 23

Considere-se, agora, um problema semelhante, porém envolvendo alunos de Física e Ciências Sociais, cuja distribuição conjunta está na Tabela 25.

Tabela 25: Distribuição conjunta das porcentagens dos 200 alunos de acordo com sexo e com o curso escolhido

Curso Escolhido	**Sexo**		**Total**
	Masculino	Feminino	
Física	100 (71%)	20 (33%)	120 (60%)
Ciências Sociais	40 (29%)	40 (67%)	80 (40%)
Total	140 (100%)	60 (100%)	200 (100%)

Fonte: Dados hipotéticos

Comparando a distribuição das porcentagens pelos cursos, independente do sexo (coluna total), com as distribuições diferenciadas por sexo (coluna de masculino e feminino), nota-se uma disparidade bem acentuada nas porcentagens. Há uma maior concentração dos homens no curso de Física e mulheres no curso de Ciências Sociais. Portanto, neste caso, parece que as variáveis: sexo e curso escolhidas estão associadas.

Pesquisa sobre consumo cultural

Será que existe algum tipo de relação entre idade de uma pessoa e o tipo de programa que ela prefere na hora de escolher entre: ir ao cinema, ir ao teatro, assistir um show de música etc.? Será que as preferências do público mais jovem são completamente diferentes das do público de meia idade? Ou será que existe um

desses programas que é sempre o preferido do público, independente da faixa etária?

Em uma pesquisa de opinião, n = 499 pessoas foram ouvidas a respeito de suas preferências em termos de consumo cultural. Admiti-se que essas pessoas representam uma amostra do "público jovem" do Rio de Janeiro. A cada um dos entrevistados perguntou-se, entre outras coisas, a sua faixa etária e qual entre cinco tipos de programa era mais do seu agrado. Com base nos resultados foi montada a seguinte Tabela 26 de contingência.

Tabela 26: Tabela de contingência relativa às variáveis: Faixa Etária e Programa Preferido em uma pesquisa de opinião sobre consumo cultural

Faixa etária	**Programa Preferido**					Total
	Cinema	Exposições	Teatro	Dança	Shows musicais	
18 a 21	68	1	15	9	45	138
22 a 25	66	3	21	12	42	144
26 a 30	66	8	24	11	25	134
31 a 40	39	3	16	8	17	83
Total	239	15	76	40	129	499

Nossa intenção é procurar extrair algumas conclusões sobre a interdependência entre "Faixa Etária" e "Programa Preferido", a partir dessa tabela de contingência. Deseja-se que essas conclusões fossem aplicáveis à população como um todo, e não apenas a essa particular amostra. Mas, neste caso, uma constatação que salta aos olhos quando se olha para a tabela de contingência é o fato de que há relativamente poucas ocorrências na coluna relativa a Exposições. Isso implica que quaisquer proporções simples que venham a ser calculadas a partir das frequências que constam nessa coluna poderão não ser estatisticamente confiáveis.

Exemplificando melhor: com base nesses dados, as pessoas que escolheram o programa Exposições se dividem pelas faixas etárias conforme a Tabela 27:

Tabela 27: Faixa etária com relação a Exposições

Faixa etária	Freq. observada	Percentagem
18 a 21	1	6,67%
22 a 25	3	20,00%
26 a 30	8	53,33%
31 a 40	3	20,00%
Total	15	100,00%

Suponha agora que dispuséssemos de uma outra amostra formada por 499 pessoas do público jovem. E que nessa outra amostra houvesse também apenas 15 pessoas optando por Exposições, porém distribuídas entre as faixas etárias de forma levemente diferente, conforme apresenta a Tabela 28.

Tabela 28: Faixa etária referentes a Exposições

Faixa etária	Freq. observada	Percentagem
18 a 21	2	13,33%
22 a 25	4	26,67%
26 a 30	7	46,67%
31 a 40	2	13,33%
Total	15	100,00%

Como pode observar, bastou introduzir uma pequena perturbação nas frequências absolutas para que ocorresse uma alteração expressiva nos percentuais. Ora, tal flutuação de uma amostra para outra é algo que está perfeitamente dentro do esperado.

Assim sendo, ficaria comprometido o nosso propósito de extrapolar para a população as conclusões extraídas a partir da amostra.

Por isso, nossa primeira providência aqui será fundir em uma só as colunas referentes à Dança e Exposições, simplesmente somando as frequências das duas. A nova coluna na criada recebe o título de Dança/Exposições. Dessa forma, a nova

tabela de contingência passou a ter quatro colunas de contagens além da coluna de totais.

Esse é um expediente muito utilizado na prática com o objetivo de se preservar a representatividade estatística dos resultados (Tabela 29).

Tabela 29: Nova tabela de contingência relativa às variáveis: Faixa Etária e Programa Preferido, após a fusão de duas colunas.

Faixa etária	**Programa preferido**				Total
	Cinema	Teatro	Shows musicais	Dança/Exposições	
18 a 21	68	15	45	10	138
22 a 25	66	21	42	15	144
26 a 30	66	24	25	19	134
31 a 40	39	16	17	11	83
Total	239	76	129	55	499

Com base na nova tabela de contingência podem ser montadas as duas tabelas de percentuais, que certamente são mais informativas sobre a eventual existência de associação entre as duas variáveis aqui consideradas.

Tabela 30: Percentuais (de linha) correspondentes aos Programas Preferidos, uma vez fixada a faixa etária

Faixa etária	**Programa Preferido**				Total (%)
	Cinema (%)	Teatro (%)	Shows musicais (%)	Dança/Exposições (%)	
18 a 21	49,28	10,87	32,61	7,25	100,00
22 a 25	45,83	14,58	29,17	10,42	100,00
26 a 30	49,25	17,91	18,66	14,18	100,00
31 a 40	46,99	19,28	20,48	13,25	100,00
Total	47,90	15,23	25,85	11,02	100,00

A Tabela 30 parece seguir, por exemplo, que:

a) Cinema é o programa preferido de praticamente metade do público considerado, independente da faixa etária.
b) Embora em todas as faixas etárias o segundo tipo de programa mais apontado seja shows musicais, há uma predominância dessa opção para o público de até 25 anos.
c) A preferência pelo teatro aumenta com a idade.

Como já visto antes, uma outra forma de encara independência entre duas variáveis “Faixa Etária” e “Programa Preferido” é inverter os papéis desempenhados por linhas e colunas, produzindo assim a tabela a seguir:

Tabela 31 Percentuais (de coluna) correspondentes às faixas etárias, uma vez fixado o programa preferido

Faixa etária	**Programa preferido**				Total
	Cinema (%)	Teatro (%)	Shows musicais (%)	Dança/Exposições (%)	
18 a 21	28,45	19,74	34,88	18,18	27,66
22 a 25	27,62	27,63	32,56	27,27	28,86
26 a 30	27,62	31,58	19,38	34,55	26,85
31 a 40	16,32	21,05	13,18	20,00	16,63
Total (%)	100,00	100,00	100,00	100,00	100,00

A Tabela 31 parece seguir, por exemplo, que praticamente 2/3 do público adepto de shows musicais estão situados nas duas primeiras faixas etárias, ou seja, têm no máximo 25 anos de idade.

Na disciplina de **estudos não paramétricos** será feita uma análise mais aprofundada das tabelas de contingência, usando o **teste qui-quadrado** para independência de variáveis.

8.4 MEDIDAS DE ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Quando existe associação entre duas variáveis, sempre é interessante quantificar essa associação. A quantificação do grau entre duas variáveis é feita através dos *coeficientes de associação ou correlação*. Essas são medias que descrevem, por meio de um único número, a dependência entre duas variáveis, no intervalo de 0 a 1, e se for próximo de zero significa falta de associação, isto é, de dependência.

Existem muitas medidas que qualificam a associação ou dependência entre duas variáveis qualitativas. Por exemplo, o *coeficiente de contingência (C),* devido a Pearson*.* Para isso, deve-se recorrer a uma importante aplicação que é o teste *qui-quadrado* ***(χ^2). Ressalta-se que esta aplicação será apresentada na disciplina de estatística não paramétrica.***

8.5 ASSOCIAÇÃO ENTRE VARIÁVEIS QUANTITATIVAS

Em muitas situações de negócios, é razoável sugerir que existam associações entre as variáveis. Por exemplo, seria lógico supor que as vendas de um item produzido em massa estejam associadas com seu preço e despesas de propaganda.

Para propósitos de tomada de decisão, é útil identificar se existe uma associação linear entre *duas variáveis* ou entre *mais de duas variáveis* e, se apropriado, quantificar a associação. Um dispositivo bastante útil para se verificar a associação entre duas variáveis quantitativas, ou entre dois conjuntos de dados, é o *diagrama de dispersão*, e sua associação pode ser quantificada utilizando-se uma medida estatística chamada *coeficiente de correlação ou grau de associação*.

Diagrama de dispersão

Um diagrama de dispersão é simplesmente uma representação de pontos de dados em um gráfico X-Y. O eixo y é utilizado para representar a variável dependente que interessa a quem toma as decisões, enquanto o eixo x é para representar uma variável que pode ser controlada ou mediada por quem toma as decisões, chamada de variável independente.

Dependendo das variáveis consideradas, a relação entre elas pode ser fortemente linear, não linear ou mesmo inexistente. Portanto, um diagrama de

dispersão é uma primeira indicação útil da possível existência de uma associação entre duas variáveis.

8.5.1 Coeficientes de associação ou correlação

A análise de correlação é uma técnica matemática utilizada para medir a força de associação entre duas variáveis. Essa medição leva em consideração a dispersão entre os valores dados. Quanto menos dispersos estiverem os dados, mais forte será a dependência, isto é, a associação entre as variáveis.

O coeficiente de correlação "**R**" assume um valor entre [– 1 e + 1] , isto é:

Se r = 1, a correlação é positiva perfeita;

Se r = -1, a correlação é negativa perfeita;

Se r = 0, a correlação é nula.

Considerando-se os dados das as variáveis X e Y, pode construir os diagramas de dispersão como mostram as Figuras 22, 23, 24 e 25.

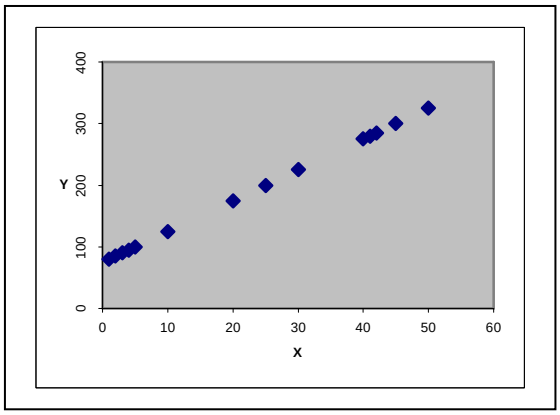


Figura 22: Associação linear positiva R = 1

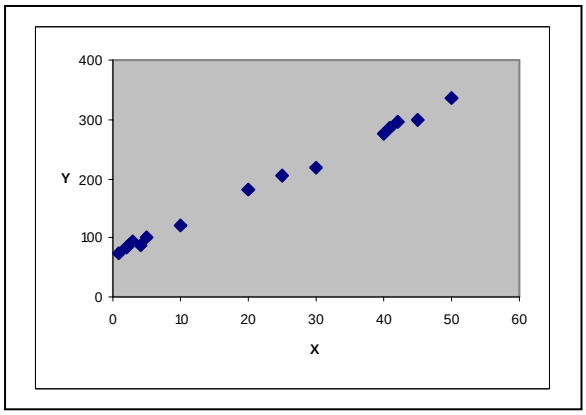


Figura 23: Associação linear positiva

Em ambas as figuras 22 e 23, nota-se que existe uma associação positiva entre as variáveis X e Y, pois à medida que aumenta uma, a outra também aumenta.

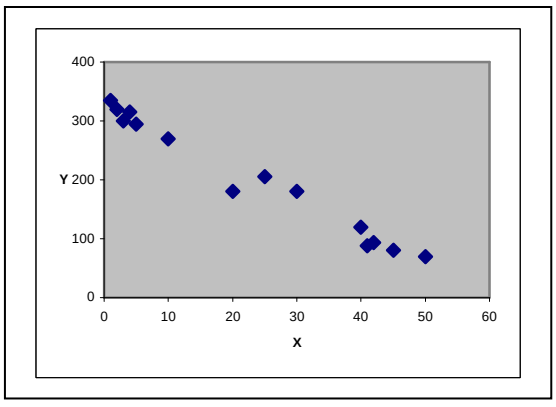


Figura 24: Associação linear negativa

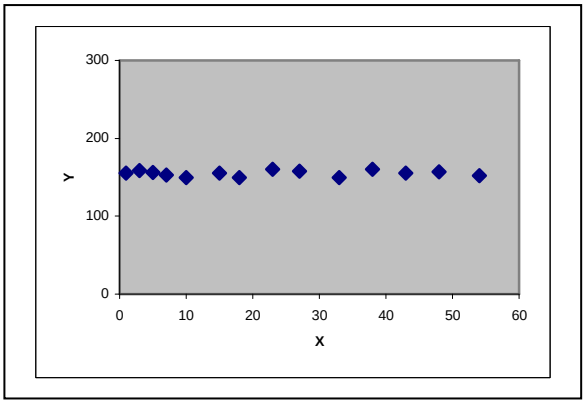


Figura 25: Não há associação - R = 0

Na figura249, existe uma associação inversa, isto é, à medida que a variável X aumenta, a variável Y diminui. Ao passo que, na figura 25 não há uma associação entre as variáveis, pois à medida que X aumenta, Y não reage.

Na Tabela 32 está apresentado os dados referentes a Taxa de Fundo de Investimento: FIC Executivo RF LP e taxa SELIC, no período de outubro de 2004 a setembro de 2006

Tabela 32: Taxa do Fundo de Investimento - FIC Executivo RF LP e taxa SELIC, no período de outubro de 2004 a setembro de 2006

Meses	**Taxa Selic (X)**	**Taxa FIC Executivo (Y)**
Out/04	1,210	1,140
Nov/04	1,250	1,190
Dez/04	1,480	1,470
Jan/05	1,380	1,336
Fev/05	1,220	1,177
Mar/05	1,530	1,485
Abri/05	1,410	1,348
Mai/05	1,500	1,430
Jun/05	1,590	1,525
Jul/05	1,510	1,429
Ago/05	1,660	1,550
Set/05	1,500	1,462
Out/05	1,410	1,347
Nov/05	1,380	1,428
Dez/05	1,470	1,460
Jan/06	1,430	1,392
Fev/06	1,150	1,098
Mar/06	1,420	1,331
Abri/06	1,080	1,002
Mai/06	1,280	1,162
Jun/06	1,180	1,097
Jul/06	1,170	1,077
Ago/06	1,260	1,153
Set/06	1,060	0,970

Fonte: Caixa Econômica Federal – 2006

Com os dados da tabela 32, constrói-se o diagrama de dispersão como mostra a Figura 26.

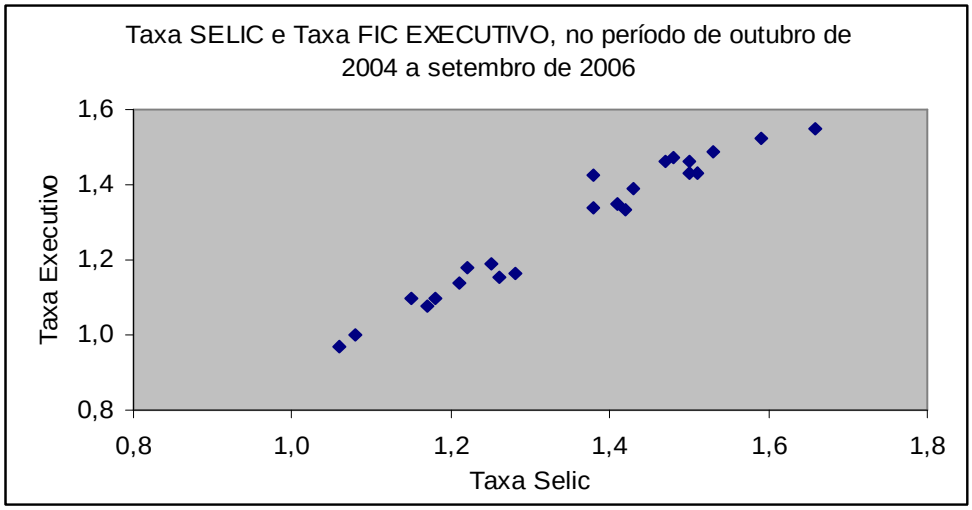


Figura 26: Diagrama de dispersão

Como já foi visto em medidas de dispersão, a soma de todos os desvios em relação à média é igual a zero, como mostra a Tabela 33.

Tabela 33: Calculo do coeficiente de correlação entre as variáveis: Selic e FIC

Meses	Selic (X)	Executivo (Y)	$X - \overline{X}$	$Y - \overline{Y}$	$Z_X = \frac{X - \overline{X}}{\sigma}$	$Z_Y = \frac{Y - \overline{Y}}{\sigma}$	$Z_X \cdot Z_y$
Out/04	1,210	1,140	-0,145	-0,154	-0,901	-0,893	0,804
Nov/04	1,250	1,190	-0,105	-0,104	-0,653	-0,603	0,394
Dez/04	1,480	1,470	0,125	0,176	0,772	1,019	0,786
Jan/05	1,380	1,336	0,025	0,042	0,152	0,243	0,037
Fev/05	1,220	1,177	-0,135	-0,117	-0,839	-0,678	0,569
Mar/05	1,530	1,485	0,175	0,191	1,082	1,106	1,196
Abri/05	1,410	1,348	0,055	0,054	0,338	0,312	0,106
Mai/05	1,500	1,430	0,145	0,136	0,896	0,787	0,705
Jun/05	1,590	1,525	0,235	0,231	1,453	1,337	1,944
Jul/05	1,510	1,429	0,155	0,135	0,958	0,781	0,748
Ago/05	1,660	1,550	0,305	0,256	1,887	1,482	2,797
Set/05	1,500	1,462	0,145	0,168	0,896	0,972	0,871
Out/05	1,410	1,347	0,055	0,053	0,338	0,306	0,104
Nov/05	1,380	1,428	0,025	0,134	0,152	0,775	0,118
Dez/05	1,470	1,460	0,115	0,166	0,710	0,961	0,682
Jan/06	1,430	1,392	0,075	0,098	0,462	0,567	0,262
Fev/06	1,150	1,098	-0,205	-0,196	-1,273	-1,136	1,446
Mar/06	1,420	1,331	0,065	0,037	0,400	0,214	0,085
Abri/06	1,080	1,002	-0,275	-0,292	-1,706	-1,692	2,887
Mai/06	1,280	1,162	-0,075	-0,132	-0,467	-0,765	0,358
Jun/06	1,180	1,097	-0,175	-0,197	-1,087	-1,142	1,241
Jul/06	1,170	1,077	-0,185	-0,217	-1,149	-1,258	1,445
Ago/06	1,260	1,153	-0,095	-0,141	-0,591	-0,817	0,483
Set/06	1,060	0,970	-0,295	-0,324	-1,830	-1,877	3,436
TOTAL	**32,530**	**31,059**	**0**	**0**			**23,504**

Fonte: Site da Caixa Econômica Federal, 2006

Observa-se que houve uma mudança de escala das colunas 2 e 3 para escala padronizada – colunas 6 e 7.

Na coluna 8 da tabela 33 ocorrem os produtos das coordenadas reduzidas. Dessa forma, pode-se definir o coeficiente de correlação entre duas variáveis X e Y como:

$$R = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right).\left(\frac{y_i - \bar{y}}{\sigma_y}\right) \qquad \text{ou} \qquad R = \frac{1}{n}\sum_{i=1}^{n}(Zx.Zy)$$

ou seja, a média dos produtos dos valores padronizados das variáveis.

Com relação ao problema anterior, tem-se: $R = \frac{1}{24}\,23{,}504 = 0{,}979$

Covariância. Uma medida de dependência linear entre duas variáveis (X, Y) é dada pela covariância:

$$Cov\,(x,\, y) = E(x\ y) - E(x).E(y)$$

onde: $E(x) = \sum_{i=1}^{n} x_i.p(x_i)$ $\quad E(y) = \sum_{i=1}^{n} y_i.p(y_i)$ $\quad E(x\ y) = \sum_{i=1}^{n} x_i.y_i\,p(x_{i,}y_i)$

Neste caso, a **Correlação linear** é dada por: $\rho(x, y) = \frac{Cov(x, y)}{\sigma_x\ .\ \sigma_y}$

Com os dados da tabela 33, vamos calcular a correlação linear entre as variáveis: Taxa Selic (X) e Taxa FIC Executivo (Y).

$$E(x) = \sum_{i=1}^{n} x_i.p(x_i) = \frac{1}{24}(1{,}210 + 1{,}250 + \cdots + 1{,}060) = 1{,}3554$$

$$E(y) = \sum_{i=1}^{n} y_i.p(y_i) = \frac{1}{24}(1{,}140 + 1{,}190 + \cdots + 0{,}970) = 1{,}2941$$

$$E(x\ y) = \sum_{i=1}^{n} x_i.y_i\,p(x_{i,}y_i) = \frac{1}{24}(1{,}379 + 1{,}487 + \cdots + 1{,}028) = 1{,}7814$$

$Cov\,(x,\, y) = E(x\ y) - E(x).E(y)$ = 1,7814 – 1,3554x 1,2941 = 0,0273

$$\rho(x,x) = \frac{Cov(x,y)}{\sigma_x \,.\, \sigma_y} = \frac{0{,}0273}{0{,}1614x0{,}1726} = 0{,}979$$

Outra maneira de se verificar se existe associação entre duas variáveis quantitativas é por meio do *coeficiente de correlação,* utilizando o método dos mínimos quadrados, tal que:

$$r = \frac{\sum x\,y - \frac{\sum x \sum y}{n}}{\sqrt{[\sum x^2 - \frac{(\sum x)^2}{n}].[\sum y^2 - \frac{(\sum y)^2}{n}]}}$$

onde $-1 \leq R \leq 1$

Esta aplicação será vista na disciplina de correlação e análise de regressão.

8.6 Associação entre as Variáveis Qualitativas e Quantitativas

É comum nessas situações analisar o que acontece com a variável quantitativa dentro de cada categoria da variável qualitativa. Essa análise pode ser conduzida por meio de medidas-resumo ou box plot.

Com os dados da Tabela 3, vamos analisar agora o comportamento dos salários dentro de cada categoria de grau de instrução, ou seja, investigar o comportamento conjunto das variáveis, *salário* e *grau de instrução*, como apresenta a Tabela 34.

Tabela 34: Medidas-resumo para a variável salário segundo o grau de instrução, na Companhia MB

SALÁRIO									
Grau de Instrução	**n**	$\bar{x}$	**σ**	**σ^2**	**X_1**	**Q_1**	**Q_2**	**Q_3**	**X_n**
Fundamental	12	7,84	2,83	8,02	4,00	6,00	7,13	9,16	13,85
Médio	18	11,53	3,61	13,04	5,73	8,83	10,91	14,42	19,40
Superior	6	16,48	4,11	16,89	10,53	13,65	16,74	18,38	23,30
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,01	23,30

Com os dados da Tabela 28 podemos construir a Figura 27 de box plot. Essa figura dá uma boa visualização e uma boa idéia para analisar a associação entre as variáveis, salário e grau de instrução.

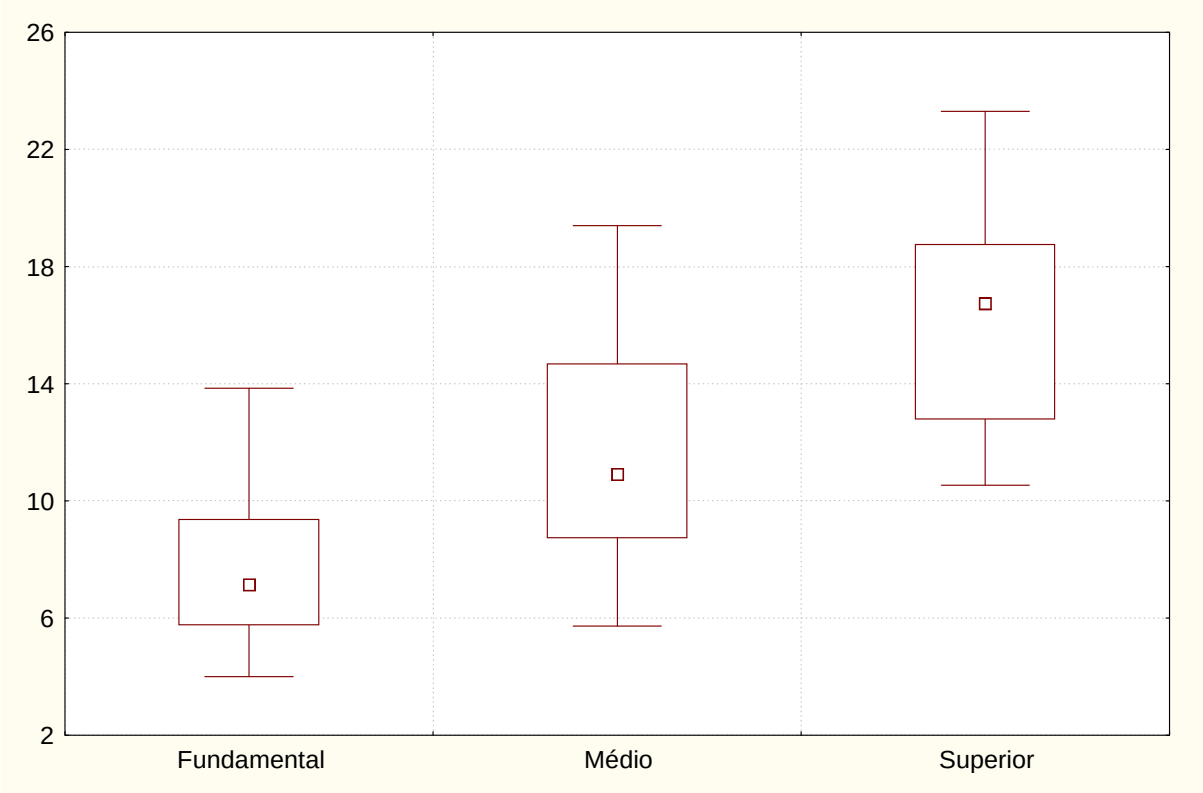


Figura 27: Salários segundo o grau de instrução dos funcionários da Companhia MB

Nota-se por meio da figura 27 uma dependência dos salários em relação ao grau de instrução: o salário aumenta conforme aumenta o nível de educação do indivíduo. O salário médio de um funcionário é 11,12 (salários mínimos), já para um

funcionário com curso superior o salário médio passa a ser 16,48, enquanto funcionários com ensino fundamental completo recebem, em média, 7,84.

Como nos casos anteriores, é interessante medir o grau de associação ou de dependência entre as duas variáveis. Com esse intuito, convém observar que as variâncias podem ser usadas como insumos para determinar essa medida. Sem usar a informação da variável categorizada, a variância calculada para a variável quantitativa para todos os dados mede a dispersão dos dados globalmente. Se a variância dentro de cada categoria for pequena e menor do que a global, significa que a variável qualitativa melhora a capacidade de previsão da quantitativa e, portanto, existe uma relação entre as duas variáveis.

Observe que, para as variáveis: *salário* e *grau de instrução*, as variâncias do salário dentro das três categorias são menores do que a global.

Neste caso, deve-se obter a variância entre as categorias da variável qualitativa, bem como a média entre elas. A média será ponderada pelo número de observações em cada categoria, ou seja;

$$\overline{\sigma^2} = \frac{\sum_{i=1}^{k} n_i . \sigma_i^{\,2}}{\sum_{i=1}^{k} n_i} = \frac{12(8{,}02) + 18(13{,}04) + 6(16{,}89)}{12 + 18 + 6} = 12{,}01$$

na qual k é o número de categorias e σ_i, a variância dos salários dentro de cada categoria i, como i = 1, 2, ..., k.

Verifica-se que $\overline{\sigma^2} \leq \sigma^2$, e o grau de associação entre as duas variáveis como ganho relativo na variância, obtido pela introdução da variável qualitativa é dado por:

$$R^2 = \frac{\sigma^2 - \overline{\sigma^2}}{\sigma^2} = 1 - \frac{\overline{\sigma^2}}{\sigma^2} \quad \rightarrow \quad R^2 = 1 - \frac{12{,}01}{20{,}46} = 0{,}413 = 41{,}3\% \quad \rightarrow \quad 0 \leq R^2 \leq 1$$

Conclui-se que 41,3% da variação total do salário é explicado pela variável grau de instrução.

Vamos analisar agora o comportamento dos *salários* dentro de cada categoria da *região procedente*, como apresenta a Tabela 35.

Tabela 35: Medidas-resumo para a variável salário, segundo a região de procedência, na Companhia MB

SALÁRIO									
Região de Procedência	**N**	$\bar{x}$	**σ**	**σ^2**	**X_1**	**Q_1**	**Q_2**	**Q_3**	**X_n**
Capital	11	11,46	5,22	27,27	4,56	7,49	9,77	16,63	19,40
Interior	12	11,55	5,07	25,71	4,00	7,81	10,65	14,70	23,30
Outra	13	10,45	3,02	9,13	5,73	8,74	9,80	13,79	16,22
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,01	23,30

Com os dados da Tabela 3 pode-se construir a Figura 28 de box plot para visualizar e analisar a associação entre as variáveis, salário e região procedência.

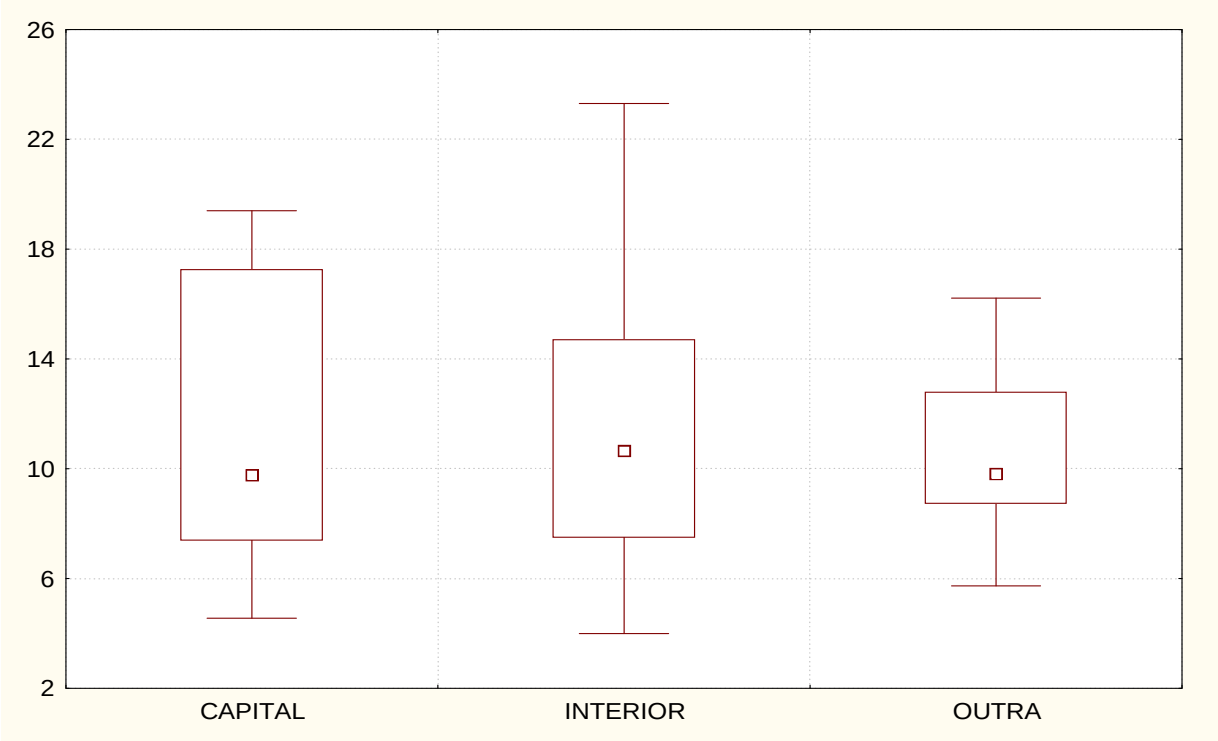


Figura 28: Salários segundo a região de procedência dos funcionários da Companhia MB

Na figura 28 temos os resultados da análise dos salários em função da região de procedência, que mostra a inexistência de uma relação melhor definida entre as duas variáveis. O salário médio de um funcionário é 11,12 (salários

mínimos), já os funcionários da capital recebem, em média, 11,46; do interior 11,55 e de outras localidades recebem, em média, 10,45.

Observe que, para as variáveis: *salário* e *região de procedência*, as variâncias do salário dentro das três categorias, ora são maiores (capital e interior) ora é menor (outros) do que a global. Neste caso, vamos calcular a variância média será ponderada pelo número de observações em cada categoria, ou seja;

$$\overline{\sigma^2} = \frac{\sum_{i=1}^{k} n_i . \sigma_i^{\,2}}{\sum_{i=1}^{k} n_i} = \frac{11(27,27) + 12(25,71) + 13(9,13)}{11 + 12 + 13} = 20,20$$

e, portanto, $R^2 = 1 - \frac{\overline{\sigma^2}}{\sigma^2}$ ➔ $R^2 = 1 - \frac{20,20}{20,46} = 0,013 = 1,3\%$

Conclui-se que apenas 1,3% da variabilidade dos salários é explicada pela região de procedência.

8.7 LISTA 3 - EXERCÍCIOS

1) Uma companhia de seguros analisou a frequência com que 2.000 segurados (1.000 homens e 1.000 mulheres) usaram o hospital. Os resultados foram:

	Homens	Mulheres
Usaram o hospital	100	150
Não usaram o hospital	900	850

a) Calcule a proporção dos homens entre os indivíduos que usaram o hospital.
b) Calcule a proporção dos homens entre os indivíduos que não usaram o hospital.
c) O uso do hospital independe do sexo do segurado?

2) Abaixo estão os dados referentes à porcentagem da população economicamente ativa empregada no setor primário e o respectivo índice de analfabetismo para algumas regiões metropolitanas brasileiras.

Regiões metropolitanas	**Setor primário (Y)**	**Índice de analfabetismo (X)**
São Paulo	2,0	17,5
Rio de Janeiro	2,5	18,5
Belém	2,9	19,5
Belo Horizonte	3,3	22,2
Salvador	4,1	26,5
Porto Alegre	4,3	16,6
Recife	7,0	36,6
Fortaleza	13,0	38,3

Fonte: Indicadores Sociais para Áreas Urbanas-IBGE-1977

a) Faça o diagrama de dispersão.
b) Você acha que existe uma dependência linear entre as duas variáveis? Se achar que sim, então calcule a correlação linear.

3) Uma pesquisa sobre a participação em atividades esportivas de adultos moradores nas proximidades de centros esportivos construídos pelo estado de São Paulo mostrou os resultados da tabela abaixo. Baseado nesses resultados você diria que a participação em atividades esportivas depende da cidade.

Participam	**Cidade**			
	São Paulo	Campinas	Rib. Preto	Santos
Sim	50	65	105	120
Não	150	185	195	180

4) Uma pesquisa para verificar a tendência dos alunos a prosseguir os estudos, segundo a classe social do respondente, mostrou a seguinte tabela:

Pretende Continuar	Classe social			Total
	Alta	Média	Baixa	
Sim	200	220	380	800
Não	200	280	720	1.200

Existe uma dependência entre os dois fatores? Por quê?

5) Completar a Tabela Medidas-resumo para a variável salário, segundo a região de procedência, na Companhia MB

SALÁRIO									
Estado Civil	**N**	$\bar{x}$	**σ**	**σ^2**	**X_1**	**Q_1**	**Q_2**	**Q_3**	**X_n**
Solteiro									
Casado									
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,01	23,30

Verifique se existe associação entre as variáveis, salário e estado civil por meio do box plot.

Calcular quanto a variação total (R^2) do salário é explicado pela variável estado civil.

REFERÊNCIAS BIBLIOGRAFIAS

ANDERSON, David R.; SWEENEY, Dennis J., WILLIAMS, Thomas A. **Estatística aplicada à administração e economia**. Trad. Luiz Sérgio de Castro Paiva. 2. ed. São Paulo: Pioneira, 2002.

BUSSAB, Wilton; MORETTIN, Pedro. **A estatística básica**. 5. ed. São Paulo: Saraiva, 2002.

MONTGOMERY, Douglas C.; RUNGER, George C., HUBELE, Norma F. **Estatística Aplicada à Engenharia**. Tradução Profa. Verônica Calado, D. Sc. 2. ed. Rio de Janeiro: LTC, 2004.

FREUND, John E.; SIMON, Gary A. **Estatística aplicada**: economia, administração e contabilidade. Trad. Alfredo Alves de Faria. 9. ed. Porto Alegre: Bookmam, 2000.

MAGALHÃES, Marcos N.; Lima. Antonio C. P. **Noções de probabilidade e estatística**. 6.ed. São Paulo: USP, 2004.

NEUFELD, John L. **Estatística aplicada à administração usando Excel**: Trad. José Luiz Celeste. São Paulo: Prentice Hall 2003.

PEREIRA, Júlio César Rodrigues. **Análise de dados qualitativos:** estratégias metodológicas para as ciências da saúde, humanas e sociais. 2.ed. São Paulo: USP, 1999.

PINHEIRO, Ismael, D. P.; CUNHA, Sonia, B. da.; CARVAJAL, Santiago, R; GOMES, Gastão, C. **Estatística básica – arte de trabalhar com dados**. Rio de Janeiro: Elsevier, 2009.

SMAILES, Joanne; McGRANE, Ângela. **Estatística aplicada à administração com excel**. São Paulo: Atlas, 2002.

SOARES, José F.; SIQUEIRA, Arminda, L. **Introdução à estatística médica.** Belo Horizonte: UFMG,1999.