INE 7001 ESTATÍSTICA PARA ADMINISTRADORES I

NOTAS DE AULA

PROF. MARCELO MENEZES REIS
MANOEL DE OLIVEIRA LINO

1 - INTRODUÇÃO

1.1 - O método científico

"A pesquisa científica é um processo de aprendizado dirigido. O objetivo dos métodos estatísticos é tornar este processo o mais eficiente possível".

O processo de pesquisa científica pode ser exemplificado através da figura abaixo:

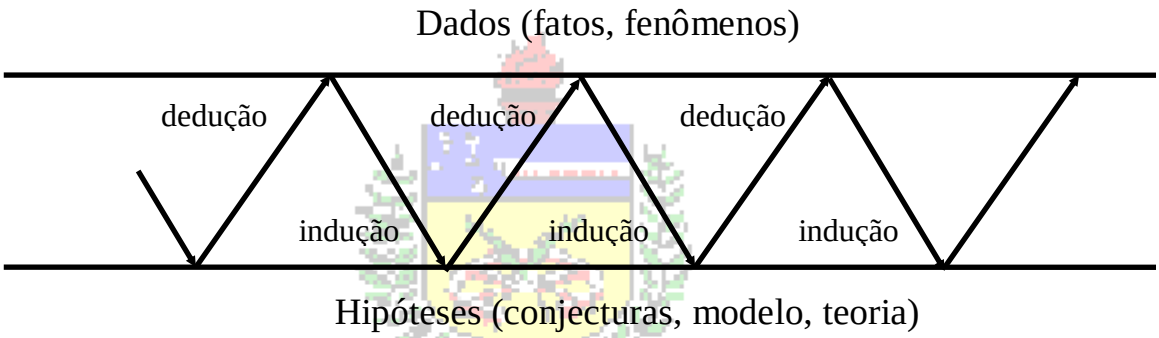


Figura 1 - Método Científico[1]

Há dados disponíveis sobre um determinado fenômeno (natural ou não) que temos interesse em COMPREENDER. Para tanto, através de um raciocínio **indutivo**, partindo do particular para o geral, procuramos montar **hipóteses** sobre o fenômeno em questão, um modelo ou teoria que explique o fenômeno.

Uma vez estabelecidas as hipóteses, estas são usadas através de um raciocínio **dedutivo**, do geral para o particular, para tentar explicar novas observações do fenômeno de interesse. Se as hipóteses conseguem explicar razoavelmente os novos dados elas são consistentes e não há necessidade de modificá-las. Mas, se os novos dados não coincidirem com o que era esperado se as hipóteses fossem verdadeiras, e os dados foram coletados corretamente, é necessário repetir todo o processo e modificar as hipóteses, ou mesmo formular novas hipóteses que expliquem aqueles dados.

Como pode ser observado na figura acima o processo é iterativo, e só há uma constante nesse processo: a MUDANÇA. O conhecimento é **mutável**, e está em permanente evolução. Os métodos estatísticos permitirão:
- garantir que os dados coletados para avaliar as hipóteses sejam válidos e representativos.
- verificar se as eventuais discrepâncias entre o que for observado e o que era esperado (sob a condição da veracidade das hipóteses) são grandes o bastante para justificar a mudança das hipóteses.

Vamos a um rápido exemplo.

Nos tempos antigos o ser humano dispunha apenas dos seus olhos para fazer observações astronômicas. Usando apenas os dados disponíveis de suas observações visuais os seres humanos de antigamente acreditavam piamente que a Terra era o centro do Universo: para eles o Sol girava em

[1] BOX, G. E. P., HUNTER, W. G. e HUNTER, J. S. - **Statistics for experimenters.** USA: John Wiley & Sons, 1978.

torno da Terra. O astrônomo egípcio de origem grega, Cláudio Ptolomeu estabeleceu no século II D.C. a teoria Geocêntrica, que explicava os movimentos dos planetas então conhecidos supondo que a Terra era o centro do Universo. A teoria Geocêntrica resistiu durante 14 séculos, em parte por considerações religiosas, mas também porque os dados disponíveis não permitiam contrariá-la. Contudo, no século XV ou XVI surgiram os primeiros telescópios óticos (foi possível o acesso a **novos** dados) e homens como Copérnico e Galileu provaram que a teoria Geocêntrica estava errada, e formularam a teoria Heliocêntrica, de que o Sol é que era o centro do Sistema Solar, pois apenas essa hipótese explicaria de forma satisfatória os novos dados que eles haviam coletado. Claro que naqueles tempos de Reforma Protestante e Contra-Reforma Católica as coisas nem sempre foram tão simples: Galileu quase foi condenado à fogueira por heresia... Mas o fato é que a teoria Heliocêntrica conseguiu explicar todos os dados que foram coletados desde então, mesmo com a descoberta de novos planetas no Sistema Solar (foram feitas algumas modificações, principalmente sobre as órbitas dos planetas que se julgava serem circulares quando na realidade são elípticas).

Muitos outros casos poderiam ser relatados, nas mais diversas áreas do conhecimento humano: a geração espontânea (em que muitos acreditaram até o século XIX), a evolução das espécies, a teoria quântica, etc.

E onde entra a Estatística nisso tudo?

1.2 - Definição de Estatística

Há dezenas de definições de Estatística, praticamente cada autor tem a sua, mas uma particularmente interessante foi apresentada pelo estatístico Paul Velleman:

"Estatística é a Ciência que permite obter conclusões a partir de dados".

É uma Ciência que parte de perguntas e desafios do mundo real:
- cientistas querem verificar se uma nova droga consegue eliminar o vírus HIV;
- uma montadora de automóveis quer verificar a qualidade de um lote inteiro de peças fornecidas através de uma pequena amostra;
- um político quer saber qual é o percentual de eleitores que votarão nele nas próximas eleições;
- os pesquisadores do departamento de aqüicultura da UFSC querem avaliar se uma nova variedade de ostra é mais produtiva do que as atualmente criadas em SC.

O principal problema que surge ao tentar responder essas perguntas é que todas as medidas feitas para tal, por mais acurados que sejam os instrumentos de medição, apresentarão SEMPRE uma variabilidade, ou seja, NÃO HÁ RESPOSTAS PERFEITAS. Feliz ou infelizmente a natureza comporta-se de forma variável: não há dois seres humanos iguais, não há dois insetos iguais, etc. Mesmo os tão comentados "clones", e os gêmeos idênticos ("clones" naturais), somente apresentam um código genético comum, se forem submetidos a experiências de vida diferentes terão um desenvolvimento distinto.

"A Estatística estuda como **controlar**, **minimizar** e **observar** a variabilidade **INEVITÁVEL** em todas as medidas e observações" feitas sobre qualquer fenômeno.

Os dados são coletados para estudar uma ou mais características de uma **POPULAÇÃO** de interesse. POPULAÇÃO é o conjunto de medidas da(s) característica(s) de interesse em todos os elementos que a(s) apresenta(m). Se, por exemplo, estamos avaliando as opiniões de eleitores sobre

os candidatos a presidente, a população da pesquisa seria constituída pelas opiniões declaradas pelos eleitores em questão[2].

Como o interesse maior está na população o ideal seria pesquisar TODA a população, em suma realizar um CENSO (como o IBGE faz periodicamente no Brasil). Contudo, por razões econômicas ou práticas (para obter rapidamente a informação ou evitar a extinção ou exaustão da população) nem sempre é possível realizar um censo, torna-se então necessário pesquisar apenas uma AMOSTRA da população, um subconjunto finito e representativo da população.

Uma das principais subdivisões da Estatística justamente é a AMOSTRAGEM, que reúne os métodos necessários para coletar adequadamente amostras representativas e suficientes para que os resultados obtidos possam ser generalizados para a população de interesse.

Após a coleta dos dados, por censo ou amostragem, a ANÁLISE EXPLORATÓRIA DE DADOS (incluindo ANÁLISE BIDIMENSIONAL, ANÁLISE DE SÉRIES TEMPORAIS E NÚMEROS ÍNDICES) permite apresentá-los e resumi-los de maneira que seja possível identificar padrões e elaborar as primeiras conclusões a respeito da população.

Posteriormente, utilizando a ESTATÍSTICA INDUTIVA (ou Inferência Estatística) é possível generalizar as conclusões dos dados para a população, e quando os dados forem provenientes de uma amostra utilizando a PROBABILIDADE para calcular a confiabilidade das conclusões obtidas[3]. Geralmente estamos interessados nos **parâmetros** (características) da população, e generalizamos os resultados das **estatísticas** da amostra (coletada para fazer tal generalização).

SEM A UTILIZAÇÃO DE MÉTODOS ESTATÍSTICOS OS RESULTADOS DE UMA PESQUISA NÃO TÊM VALIDADE CIENTÍFICA.

Vamos ver então os dois tipos básicos de pesquisa estatística.

1.3 - Tipos de Pesquisa Estatística

A pesquisa estatística pode ser conduzida basicamente de duas formas, independente de ser por censo ou amostragem: Levantamento e Experimento.

No Levantamento, também chamado de Pesquisa Correlacional, são apenas observadas as características da população, com pouco ou nenhum controle por parte do pesquisador. Esse tipo de pesquisa costuma gerar grandes volumes de dados, mas não é possível provar relações de causa e efeito com um Levantamento (nem todas as causas de variação estão sob controle), apenas afirmar que devem existir relações entre as variáveis sob análise. Como exemplo de Levantamento: o Censo do IBGE, pesquisas de opinião pública, etc.

[2] É muito comum definir População como sendo um conjunto de elementos com pelo menos uma característica em comum, assim no exemplo a população seria formada pelos pacientes que apresentam a característica em comum, a suspeita de diabetes. Embora mais simples, essa definição não é totalmente correta, pois o interesse maior está nas medidas do nível de glicose que permitirão classificar os pacientes como diabéticos ou não.

[3] Quando toda a população é pesquisada por CENSO (corretamente executado) teoricamente não há incerteza, portanto não há necessidade de calcular a confiabilidade das conclusões obtidas, e o estudo resume-se à Análise Exploratória de Dados. Contudo, o censo mais bem conduzido não impede que sejam cometidos erros de medição ou que os respondentes mintam ou omitam dados importantes.

O segundo tipo de pesquisa é o Experimento, ou Pesquisa Experimental. Neste caso o pesquisador tem um grande controle sobre as condições de pesquisa, praticamente eliminado todas as fontes "indesejáveis" de variação através de um PLANEJAMENTO DO EXPERIMENTO. Sendo assim é o único tipo de pesquisa que permite provar conclusivamente relações de causa e efeito. Devido ao maior controle sobre as causas de variação não há necessidade de gerar um volume de dados tão grande como no caso do Levantamento. O Experimento é largamente empregado em farmacologia (para testar a eficácia de novos remédios e vacinas), e no ambiente industrial, e em todas as situações em que é necessário provar relações de causa e efeito e seja possível controlar as causas de variação.

No Experimento é muito comum testar se dois ou mais "tratamentos" em amostras representativas da população:

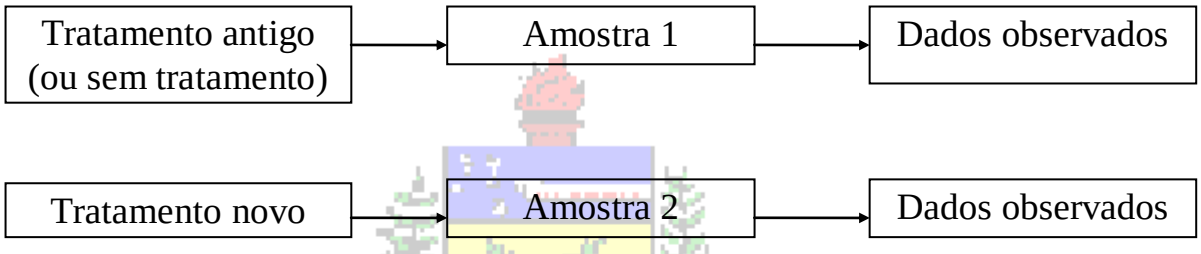


Figura 2 - Experimento

1.4 - Arredondamento Estatístico

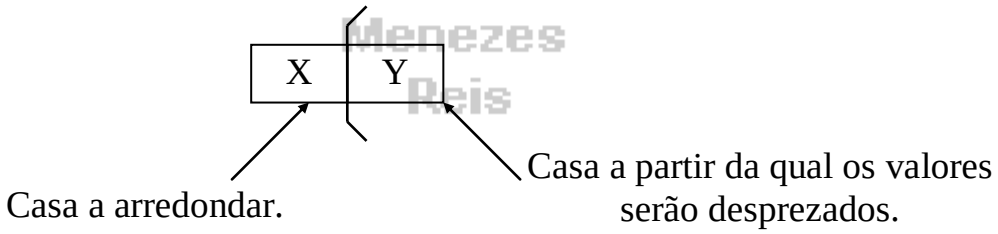


Figura 3 - Casas envolvidas em arredondamento estatístico

- Se Y estiver entre 0 e 4 deixar X como está.
- Se Y estiver entre 6 e 9, X deve ser acrescido de uma unidade.
- Se Y for igual a 5:
    - Se após Y houver outros números (diferentes de zero), X deve ser acrescido de uma unidade.
    - Se após Y não houver números diferentes de zero:
        - Se X for par, deve ser deixado como está.
        - Se X for ímpar, deve ser acrescido de uma unidade.

Arredondar	para	resultado
1,1213	milésimo	1,121
6,586	centésimo	6,59
12,57585	centésimo	12,58
6,23515	centésimo	6,24
9,65	décimo	9,6
9,75	décimo	9,8

1.5 - Estrutura de um arquivo de dados

Nos Capítulos 6 e 7 aprenderemos como obter os dados para a realização de análises, estudando o Planejamento de Pesquisa e Técnicas de Amostragem. Contudo, uma vez disponíveis os dados precisam ser tabulados, para possibilitar sua análise. Atualmente os dados costumam ser armazenados em meio computacional: seja em grandes bases de dados, programas estatísticos ou mesmo planilhas eletrônicas. Quer sejam oriundos de pesquisa de campo, ou apenas registros de operações financeiras, arquivos de recursos humanos, entre outros. Universalmente os dados costumam ser armazenados em uma estrutura fixa, que possibilita a aplicação de várias técnicas para extrair as informações de interesse.

As variáveis são registradas nas colunas, e os casos nas linhas. As variáveis são as características pesquisadas ou registradas. Imagine a base de dados do DAE da UFSC, que armazena as informações dos acadêmicos, contendo as variáveis nome do aluno, data de nascimento, número de matrícula, IAA, IAP, e outras informações, ou uma operadora de cartão de crédito, que armazena as transações efetuadas, contendo o número do cartão, nome do titular, hora da transação, valor do crédito, bem ou serviço adquirido. Os casos constituem cada indivíduo ou registro, para a base do DAE, João Ninguém, nasceu em 20 de fevereiro de 1985, matrícula 02xxxxxxx-01, IAA = 3,5, IAP = 6,0. Para a operadora de cartão de crédito, cartão número xxxxxxxxx-84, José Nenhum, R$200, 14h28min - 11 de dezembro de 2016, supermercado.

Imagine uma organização que registre os dados dos seus funcionários, como sexo, idade, anos de educação, função, salário anual, salário inicial, tempo de serviço na organização, experiência prévia, e nacionalidade. Veja a Figura 4.

VALU LABE	1 SEXO	2 IDADE	3 ANOSEDUC	4 FUNCAO	5 SALARIOA	6 SALARIOI	7 ANOSSERV	8 EXPERPR	9 NACIONAL
1	Masculino	49.	15	Gerência	57000.00	27000.00	8.17	12.00	Brasileiro
2	Masculino	43.	16	Escritório	40200.00	18750.00	8.17	3.00	Brasileiro
3	Feminino	71.	12	Escritório	21450.00	12000.00	8.17	31.75	Brasileiro
4	Feminino	54.	8	Escritório	21900.00	13200.00	8.17	15.83	Brasileiro
5	Masculino	46.	15	Escritório	45000.00	21000.00	8.17	11.50	Brasileiro
6	Masculino	42.	15	Escritório	32100.00	13500.00	8.17	5.58	Brasileiro
7	Masculino	45.	15	Escritório	36000.00	18750.00	8.17	9.50	Brasileiro
8	Feminino	35.	12	Escritório	21900.00	9750.00	8.17	0.00	Brasileiro
9	Feminino	55.	15	Escritório	27900.00	12750.00	8.17	9.58	Brasileiro
10	Feminino	55.	12	Escritório	24000.00	13500.00	8.17	20.33	Brasileiro
11	Feminino	51.	16	Escritório	30300.00	16500.00	8.17	11.92	Brasileiro
12	Masculino	35.	8	Escritório	28350.00	12000.00	8.17	2.17	Estrangeiro
13	Masculino	40.	15	Escritório	27750.00	14250.00	8.17	2.83	Estrangeiro
14	Feminino	52.	15	Escritório	35100.00	16800.00	8.17	11.42	Estrangeiro
15	Masculino	38.	12	Escritório	27300.00	13500.00	8.08	5.50	Brasileiro
16	Masculino	36.	12	Escritório	40800.00	15000.00	8.08	2.00	Brasileiro
17	Masculino	38.	15	Escritório	46000.00	14250.00	8.08	4.00	Brasileiro
18	Masculino	45.	16	Gerência	103750.0	27510.00	8.08	5.83	Brasileiro

Figura 4 - Dados dos funcionários

Veja que cada uma das variáveis é registrada em uma coluna específica, e que nas linhas encontram-se os registros de cada funcionário. Por exemplo, o funcionário 1 é do sexo masculino, tem 49 anos, 15 anos de educação, exerce função de gerência, ganha 57000 ao ano, iniciou na empresa ganhando 27000, tem 8,17 anos de serviço, 12 anos de experiência prévia e é brasileiro.

A esmagadora maioria dos programas estatísticos, gerenciadores de bases de dados e planilhas eletrônicas com capacidade estatística exige que os dados sejam estruturados de acordo com o formato da Figura 4. Podem-se ter tantas colunas e linhas quantas se quiser, respeitando, porém as capacidades dos programas, o Microsoft Excel 2003®, por exemplo, admite apenas 65000 linhas, o que é o suficiente para muitas aplicações.

2 - ANÁLISE EXPLORATÓRIA DE DADOS

A Análise Exploratória de Dados, antigamente chamada apenas de Estatística Descritiva, constitui o que a maioria das pessoas entende como Estatística, e inconscientemente usa no dia a dia. Consiste em RESUMIR E ORGANIZAR os dados coletados através de tabelas, gráficos ou medidas numéricas, e a partir dos dados resumidos procurar alguma regularidade ou padrão nas observações (INTERPRETAR os dados).

A partir dessa interpretação inicial é possível identificar se os dados seguem algum modelo conhecido, que permita estudar o fenômeno sob análise, ou se é necessário sugerir um novo modelo.

2.1 - Variáveis

Quando um determinado fenômeno é estudado determinadas características são analisadas: as **variáveis**. É através das variáveis que se torna possível descrever o fenômeno. As variáveis são características que podem ser observadas ou medidas em cada elemento pesquisado (seja por censo ou amostragem, levantamento ou experimento), sob as mesmas condições. Para cada variável, para cada elemento pesquisado, em um dado momento, há um e apenas um resultado possível.

As variáveis podem basicamente ser classificadas de acordo com o seu **nível de mensuração** (o quanto de informação cada variável apresenta) e seu **nível de manipulação** (como uma variável relaciona-se com as outras no estudo), Veja a Figura 5 e Figura 6.

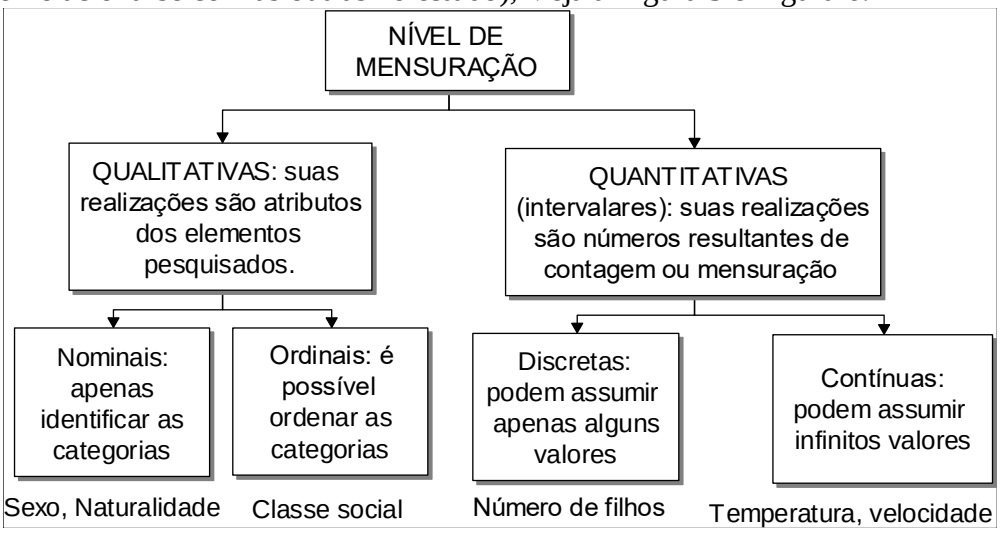


Figura 5 - Classificação das variáveis por nível de mensuração

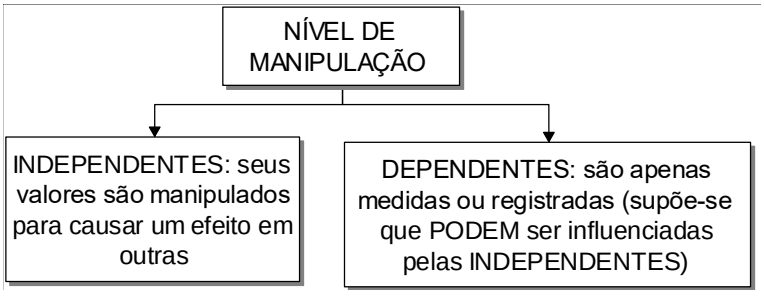


Figura 6 - Classificação das variáveis por nível de manipulação

2.1.1 - Classificação por nível de mensuração

A primeira classificação divide as variáveis em QUALITATIVAS e QUANTITATIVAS. As variáveis QUALITATIVAS ou categóricas são aquelas cujas realizações são atributos (categorias) do elemento pesquisado, como o sexo, grau de instrução, espécie. As variáveis QUALITATIVAS podem ser NOMINAIS ou ORDINAIS.

As variáveis NOMINAIS podem ser medidas apenas em termos de quais itens pertencem a diferentes categorias, mas não se pode quantificar nem mesmo ordenar tais categorias. Por exemplo, pode-se dizer que 2 indivíduos são diferentes em termos da variável A (sexo, por exemplo), mas não se pode dizer qual deles "tem mais" da qualidade representada pela variável. Exemplos típicos de variáveis nominais são sexo, naturalidade, etc. As variáveis ORDINAIS permitem ordenar os itens medidos em termos de qual tem menos e qual tem mais da qualidade representada pela variável, mas ainda não permitem que se diga "o quanto mais". Um exemplo típico de uma variável ordinal é o status sócio-econômico das famílias residentes em uma localidade: sabe-se que média-alta é mais "alta" do que média, mas não se pode dizer, por exemplo, que é 18% mais alta. A própria distinção entre mensuração nominal, ordinal e intervalar representa um bom exemplo de uma variável ordinal: pode-se dizer que uma medida nominal provê menos informação do que uma medida ordinal, mas não se pode dizer "quanto menos" ou como esta diferença se compara à diferença entre mensuração ordinal e quantitativa.

Já as variáveis QUANTITATIVAS são aquelas cujas realizações são números resultantes de contagem ou mensuração, como número de filhos, número de clientes, velocidade em km/h, peso em kg, etc. As variáveis quantitativas também costumam ser divididas em DISCRETAS e CONTÍNUAS.

As variáveis QUANTITATIVAS DISCRETAS são aquelas que podem assumir apenas alguns valores numéricos que geralmente podem ser listados (número de filhos, número de acidentes). As VARIÁVEIS QUANTITATIVAS CONTÍNUAS são aquelas que podem assumir teoricamente qualquer valor em um intervalo (velocidade, peso). A predileção dos pesquisadores em geral por variáveis quantitativas explica-se porque elas costumam conter mais informação do que as qualitativas. Quando a variável peso de um indivíduo é descrita em termos de "magro" e "gordo" sabemos que o gordo é mais pesado do que o magro, mas não temos idéia de quão mais pesado. Se, contudo, descreve-se o peso de forma numérica, medido em quilogramas, e um indivíduo pesa 60 kg e outro pesa 90 kg, não somente sabemos que o segundo é mais pesado, mas que é 30 kg mais pesado do que o primeiro[4].

É importante ressaltar que a forma como a variável está sendo medida definirá o seu nível de mensuração. Por exemplo, a variável velocidade de um carro. Se definirmos velocidade como resultado de uma medição por meio de radar resultando em um valor em km/h trata-se de uma variável quantitativa contínua. Se, porém, definirmos a velocidade como resultado de uma medição em que alguém declara a velocidade como "baixa", "média" ou "alta", ela passa a ser qualitativa ordinal.

2.1.2 - Classificação pelo nível de manipulação

Outra forma de classificar as variáveis refere-se à sua manipulação: variáveis INDEPENDENTES e DEPENDENTES.

[4] Nem sempre, porém uma variável pode ser descrita em termos quantitativos, muitas pesquisas foram prejudicadas pela tentativa de quantificar todas as variáveis envolvidas, por exemplo, inteligência e criatividade.

Variáveis INDEPENDENTES são aquelas que são manipuladas enquanto que variáveis DEPENDENTES são apenas medidas ou registradas (como resultado da manipulação das variáveis independentes). Esta distinção confunde muitas pessoas que dizem que "todas as variáveis dependem de alguma coisa". Entretanto, uma vez que se esteja acostumado a esta distinção ela se torna indispensável.

"As variáveis independentes são aquelas que PODEM INFLUENCIAR os valores das variáveis dependentes". Somente a realização do estudo vai permitir verificar se há realmente tal influência: e somente poderemos afirmar que a variável independente é a CAUSA da variável dependente assumir determinado resultado se o estudo for um experimento (pesquisa experimental).

Os termos variável dependente e independente aplicam-se principalmente à pesquisa experimental, onde algumas variáveis são manipuladas, e, neste sentido, são "independentes" dos padrões de reação inicial, intenções e características das unidades experimentais. Espera-se que outras variáveis sejam "dependentes" da manipulação ou das condições experimentais. Ou seja, elas dependem "do que as unidades experimentais farão" em resposta. Contrariando um pouco a natureza da distinção, esses termos também são usados em estudos em que não se manipulam variáveis independentes, literalmente falando, mas apenas se designam sujeitos a "grupos experimentais" (blocos) baseados em propriedades pré-existentes dos próprios sujeitos.

Exemplo 2.1 - A empresa Escolástica Ltda. quer conhecer o perfil dos seus 474 funcionários[5], para responder às seguintes perguntas:
- Identificar se há predominância masculina ou feminina.
- Mensurar a qualificação do pessoal (pelos anos de escolaridade).
- Verificar se como está o "turnover": avaliando as idades, tempo de serviço e experiência prévia.

Para tanto, dispõe da seguinte base de dados, parcialmente mostrada na Figura 7.

VALU LABE	1 SEXO	2 IDADE	3 ANOSEDUC	4 FUNCAO	5 SALARIOA	6 SALARIOI	7 ANOSSERV	8 EXPERPR	9 NACIONAL
1	Masculino	49.	15	Gerência	57000.00	27000.00	8.17	12.00	Brasileiro
2	Masculino	43.	16	Escritório	40200.00	18750.00	8.17	3.00	Brasileiro
3	Feminino	71.	12	Escritório	21450.00	12000.00	8.17	31.75	Brasileiro
4	Feminino	54.	8	Escritório	21900.00	13200.00	8.17	15.83	Brasileiro
5	Masculino	46.	15	Escritório	45000.00	21000.00	8.17	11.50	Brasileiro
6	Masculino	42.	15	Escritório	32100.00	13500.00	8.17	5.58	Brasileiro
7	Masculino	45.	15	Escritório	36000.00	18750.00	8.17	9.50	Brasileiro
8	Feminino	35.	12	Escritório	21900.00	9750.00	8.17	0.00	Brasileiro
9	Feminino	55.	15	Escritório	27900.00	12750.00	8.17	9.58	Brasileiro
10	Feminino	55.	12	Escritório	24000.00	13500.00	8.17	20.33	Brasileiro
11	Feminino	51.	16	Escritório	30300.00	16500.00	8.17	11.92	Brasileiro
12	Masculino	35.	8	Escritório	28350.00	12000.00	8.17	2.17	Estrangeiro
13	Masculino	40.	15	Escritório	27750.00	14250.00	8.17	2.83	Estrangeiro
14	Feminino	52.	15	Escritório	35100.00	16800.00	8.17	11.42	Estrangeiro
15	Masculino	38.	12	Escritório	27300.00	13500.00	8.08	5.50	Brasileiro
16	Masculino	36.	12	Escritório	40800.00	15000.00	8.08	2.00	Brasileiro
17	Masculino	38.	15	Escritório	46000.00	14250.00	8.08	4.00	Brasileiro
18	Masculino	45.	16	Gerência	103750.0	27510.00	8.08	5.83	Brasileiro

Figura 7 - Dados dos funcionários da empresa Escolástica Ltda. (parcial)

a) Identificar os níveis de mensuração das 9 variáveis?
Sexo, Idade em anos completos, Anos de educação completos (ANOSEDUC), Função (serviços gerais, escritório, gerência), Salário atual anual em reais (SALARIOA), Salário Inicial anual em reais (SALARIOI), Anos de serviço em anos (ANOSSERV), Experiência prévia em anos (EXPERPR), Nacionalidade (NACIONAL).
b) Há interesse em obter sumários descrevendo:
- as funções exercidas de acordo com o sexo do funcionário.
- os salários atuais em função do sexo do funcionário.

[5] O arquivo de dados (no formato do Microsoft Excel ® está disponível no ambiente moodle ou na minha página pessoal: http://www.inf.ufsc.br/~marcelo.menezes.reis/INE7001.html

- os salários atuais em função dos anos de educação do funcionário.

Quais são as variáveis independente e dependente em cada caso?

a) Para identificar os níveis de mensuração é preciso avaliar como a variável está sendo medida. Com isso teremos:

- sexo e nacionalidade: apresentam duas categorias (masculino e feminino; brasileiro e estrangeiro); são variáveis qualitativas (pois suas realizações são atributos) nominais (porque não é possível ordenar seus atributos, masculino não é "menos" do que feminino e vice-versa).

- idade e anos de educação: medidas em anos completos (observe na figura 6 que não há valores decimais); são variáveis quantitativas (pois suas realizações são números) discretas (porque somente podem assumir alguns valores[6]).

- salário atual, salário inicial, anos de serviço e experiência prévia: podem assumir uma infinidade de valores (por serem medidas em milhares, como os salários, ou por permitirem a existência de valores decimais); são variáveis quantitativas (pois suas realizações são números) contínuas (porque podem assumir muitos valores).

- função: apresenta três categorias; é uma variável qualitativa (pois suas realizações são atributos) ordinal (pois podemos estabelecer uma ordenação dos seus atributos, quem exerce função de gerência é "mais importante" do que quem exerce função de escritório, ao menos no que tange à tomada de decisões).

b) Para identificar as variáveis independente e dependente devemos observar o objetivo de cada relatório.

- Relatório da função de acordo com o sexo do funcionário. O sexo "poderia influenciar" a função exercida (ou mais rigorosamente, haveria alguma associação[7] entre a função exercida e o sexo do funcionário). Então sexo seria a variável independente e função seria a dependente.

- Relatório do salário atual em função do sexo do funcionário. Novamente, o sexo "poderia influenciar" o valor do salário atual (ou mais rigorosamente, haveria alguma associação entre o salário atual e o sexo do funcionário). Então sexo seria a variável independente e salário atual seria a dependente.

- Relatório do salário atual em função dos anos de educação. Os anos de educação "poderiam influenciar" no salário atual (ou mais rigorosamente, haveria alguma correlação[8] entre o salário atual e o número de anos de educação do funcionário). Então anos de educação seria a variável independente e salário atual a dependente.

2.2 - Pré-análise dos Dados

Todas as técnicas estatísticas apresentadas na literatura, para variáveis qualitativas ou quantitativas, tem duas suposições básicas:

1) Os dados obtidos são confiáveis, o planejamento da pesquisa garante que eles permitirão responder as perguntas formuladas inicialmente.

2) A base de dados, provavelmente armazenada em computador está “limpa”, podendo ser analisada imediatamente.

Infelizmente, em muitas situações reais nenhuma das duas suposições é satisfeita... A primeira é obviamente mais importante, mas vamos nos ater a segunda neste texto. Antes de aplicar técnicas estatísticas como tabelas, gráficos e medidas de síntese, é preciso realizar uma **pré-análise dos dados**.

[6] Sem esquecer, porém, que pode haver uma grande disparidade entre as idades, o que, talvez, configuraria uma variável contínua.

[7] Quando estamos avaliando o relacionamento entre duas variáveis qualitativas, ou entre uma qualitativa e uma quantitativa, chamamos tal relacionamento de associação.

[8] Quando estamos avaliando o relacionamento entre duas variáveis quantitativas, tal relacionamento é chamado correlação.

Usando ferramentas computacionais de filtragem, disponíveis em planilhas eletrônicas e programas estatísticos, é possível identificar os diferentes valores que as variáveis qualitativas e quantitativas podem assumir. O objetivo é identificar os **dados perdidos**, **erros de registro**, **valores discrepantes** e **inconsistências**. A existência de tais problemas pode afetar seriamente as conclusões porventura obtidas a partir dos dados, e, portanto pôr em risco a qualidade das decisões decorrentes.

2.2.1 - Dados perdidos (missing data)

Dados perdidos são valores de uma variável que não estão disponíveis no conjunto de dados, estão "em branco". Por exemplo, em uma pesquisa de opinião eleitoral algumas pessoas podem não declarar seu voto, resultando em dados perdidos, ou o famoso "não respondeu". Muitos estatísticos afirmam que é virtualmente impossível obter um conjunto de dados sem dados perdidos, especialmente aqueles oriundos de pesquisas de opinião (eleitoral, de mercado, socioeconômica) e mesmo pesquisas médicas. Aceita-se até cerca de 5% de dados perdidos em uma base de dados, mais do que isso o processo de pesquisa, e/ou de registro dos dados, pode ter sido prejudicado de tal forma que os resultados podem não ser confiáveis.

Os dados perdidos podem ter várias causas. No caso das pesquisas de opinião, as questões (ou as opções de resposta) podem não ser compreendidas pelos respondentes, que preferem não responder, ou nenhuma das opções contempla sua verdadeira opinião. Ou ainda, as questões (ou as opções) foram omitidas pelos entrevistadores, deliberadamente ou não, o que fez com que não houvesse respostas... Há ainda os casos em que os respondentes, embora compreendendo as questões e opções resolve não expressar sua opinião (seja por medo de represálias, especialmente em pesquisas de opinião eleitoral ou em ambientes de trabalho, ou por querer resguardar sua privacidade). Em estudos com animais, plantas, e mesmo seres humanos, os dados perdidos costumam também ocorrer devido à morte dos sujeitos, ou especialmente no acompanhamento de doenças em seres humanos, porque o sujeito decide não mais fornecer os dados aos pesquisadores[9]. É possível também que simplesmente alguém se esqueceu de registrar os dados, pelas mais diversas razões. Veja um conjunto com dados perdidos:

A	B	C	D	E	F	G	H
108	Masculino	TV	12.1	Abaixo da concorrência	Ótima	Regular	100
109	Feminino	Degustação	4.85		Regular	Bom	4
110	Feminino	Degustação	10.87	Acima da concorrência	Regular	Bom	8
111	Masculino	Rádio	12.58	Semelhante à concorrência	Regular	Regular	5
112	Feminino	TV	2.2	Acima da concorrência	Péssima	Bom	8
113	Masculino	Rádio	8.09	Semelhante à concorrência	Boa	Bom	2
114	Masculino	Degustação	11.42	Semelhante à concorrência	Ótima	Ótimo	1
115	Feminino	TV	13.07	Semelhante à concorrência	Boa	Regular	3
116	Masculino	TV	16.94	Abaixo da concorrência	Ótima	Bom	2
117	Masculino	TV	8.67	Acima da concorrência	Ruim	Bom	6
118	Feminino	Rádio	10.44	Semelhante à concorrência	Boa	Regular	5
119	Masculino	Rádio	6.61	Acima da concorrência	Ruim	Bom	4
120	Masculino	TV	14.33	Acima da concorrência	Boa	Bom	4
121	Feminino	TV	3.68	Acima da concorrência	Ruim	Bom	6
122	Feminino	TV	8.01	Acima da concorrência	Regular	Bom	6
123	Feminino	TV	7.19	Semelhante à concorrência	Ruim	Bom	6
124	Masculino	Rádio	8.25	Semelhante à concorrência	Boa	Ruim	0.4
125	Feminino	Rádio	4.74	Semelhante à concorrência	Boa	Regular	2
126	Masculino	Rádio	20.56	Abaixo da concorrência	Boa	Ruim	6
127	Masculino	Rádio	9.87	Abaixo da concorrência	Ótima	Ruim	2
128	Feminino	Rádio	12.53	Semelhante à concorrência	Ótima	Bom	4
129	Masculino	TV	6.79	Semelhante à concorrência	Boa	Bom	3
130	Masculino	Rádio	12.85	Semelhante à concorrência	Boa	Bom	4
131	Masculino	TV		Abaixo da concorrência	Ótima	Bom	3
132	Feminino	TV	14.99	Abaixo da concorrência	Ótima	Bom	1

Na Figura 8 podemos ver que algumas células estão vazias, na coluna D, linha 131, o valor de renda não foi registrado, e na coluna E, linha 109 uma opinião sobre preço de um produto também não.

O que fazer nestas circunstâncias? Há basicamente cinco cursos de ação possíveis: a eliminação completa do registro, a eliminação parcial do registro, o preenchimento com base na média da variável, o preenchimento por interpolação, ou a criação da categoria "não respondeu".

Figura 8 - Conjunto com dados perdidos

[9] Muitas pessoas durante estudos médicos, envolvendo alguma espécie de terapia, passam a sentirem-se melhor, e decidem não mais comparecer às visitas periódicas de acompanhamento, que podem se estender por muito tempo, causando a existência de dados perdidos.

A eliminação completa ("casewise deletion") consiste em simplesmente eliminar todos os casos (linhas) que apresentem pelo menos um dado perdido. Para o arquivo da Figura 8 - Conjunto com dados perdidos, seriam eliminados os registros (linhas) 109 e 131. Realiza-se então a análise dos outros dados. Já a eliminação parcial ("pairwise deletion") elimina os casos apenas nas operações que envolvem as variáveis que apresentam dados perdidos. Para o caso da Figura 8 - Conjunto com dados perdidos, nas operações envolvendo as colunas D e E os dados das linhas 131 e 109 (respectivamente) não seriam computados, mas não seriam removidos do arquivo. Tanto a eliminação completa quanto a parcial estão disponíveis em muitos programas estatísticos, mas em planilhas eletrônicas a implementação da parcial é mais complexa. Ambas apresentam também o inconveniente, maior na eliminação total, de causar **perda de informação**.

O preenchimento dos valores perdidos permite mitigar o efeito da perda de informação. Há duas possibilidades. Em uma delas no lugar dos dados perdidos são postos os valores da média[10] da variável, supondo que a média seja representativa dos valores que a variável pode assumir. Na outra possibilidade utiliza-se uma interpolação para estimar os dados perdidos: projeta-se uma curva dos dados para inferir os perdidos (o que exige o estudo do relacionamento com outras variáveis). Ambas as possibilidades são mais indicadas para variáveis *quantitativas*, pois para variáveis qualitativas calcular médias e realizar interpolações não faz sentido, mesmo que os dados tenham sido codificados numericamente eles são intrinsecamente qualitativos. O grande problema do preenchimento dos valores perdidos é a **criação de informação**, mesmo que a média seja um bom representante dos valores da variável, ou a interpolação seja acurada.

Outra solução para os dados perdidos é simplesmente aceitar a sua existência. Para uma variável qualitativa considera-se que os dados perdidos constituem mais um dos valores, por exemplo, "não respondeu" (bastante comum em pesquisas de opinião) ou "não disponível", e prossegue-se com a análise dos dados. Na análise de variáveis quantitativas a maioria dos programas estatísticos e planilhas eletrônicas desconsideram os valores perdidos ao calcular as medidas de síntese, e ao construir distribuições de frequência pode-se colocar como nota de rodapé a quantidade de dados perdidos encontrada.

2.2.2 – Erros de registro

Os erros de registro são valores que foram armazenados incorretamente na base de dados, geralmente são erros grosseiros, fáceis de identificar e corrigir.

Nas variáveis qualitativas os erros de registro costumam ser resultado da falta de uniformidade no armazenamento dos valores. Por exemplo, imagine a variável qualitativa "turno", que poderia assumir os valores Matutino, Vespertino e Noturno: algum digitador descuidado poderia registrar Mat, ou Matuti ao invés de Matutino, o que cria novos valores para a variável turno. Erros ortográficos (por exemplo, Maututino, ou Mattutino) também costumam ser fonte de erros de registro. A identificação dos erros pode ser feita através da construção de uma distribuição de frequências (preferencialmente através de uma ferramenta computacional), que relacionará os diferentes valores que a variável apresenta no conjunto de dados. Para corrigir os erros de registro basta varrer a base de dados, geralmente usando uma ferramenta de substituição (disponível em praticamente todos os programas estatísticos, planilhas eletrônicas e gerenciadores de bases de dados) para uniformizar os valores.

Nas variáveis quantitativas é necessário cuidado para não confundir erros de registro com valores discrepantes. Os erros seriam valores "impossíveis" para a variável, por exemplo, altura e peso de uma pessoa com valores negativos (...), ou alguma criança em ensino pré-escolar que

[10] Maiores detalhes sobre como calcular a média na seção 2.4.

apresente idade igual a 400 anos (admite-se que seja 4 anos)... É preciso um exame cuidadoso para evitar a confusão entre valor discrepante (por exemplo, uma renda de 200 salários mínimos) com erro de registro (por exemplo, uma renda de –200 salários mínimos).

2.2.3 – Valores discrepantes

Mais aplicável às variáveis *quantitativas*. Valores discrepantes são aqueles que estão muito acima, ou muito abaixo da maioria dos valores do conjunto de dados. Por exemplo, houve um contribuinte no Brasil que em certo ano chegou a pagar 63 milhões de reais de imposto de renda... Se for descartada a hipótese de erro de registro (ver seção 2.2.2) os valores discrepantes devem ter uma atenção especial, pois podem indicar situações inesperadas[11].

Imagine que a variável Renda (em salários mínimos) está sendo avaliada em um grupo de 5000 pessoas. A maioria apresenta renda de 1 a 8 salários mínimos, e alguns poucos apresentam valores de 25, 30 e 40 salários mínimos – valores discrepantes superiores. Outro caso seria a variável Receita Mensal (em reais) de uma rede de lojas: a maioria apresenta valores em torno de 500 ou 600 mil reais, e surgem lojas com 10 mil ou 20 mil reais – discrepantes inferiores.

A identificação de valores discrepantes pode ser feita através de distribuições de frequências (agrupadas em classes ou não), e pela identificação de valores máximos e mínimos das variáveis. Na seção 2.6 aprenderemos métodos numéricos para identificar valores discrepantes.

2.2.4 – Inconsistências

As inconsistências nos conjuntos de dados nem sempre são fáceis de identificar. Por exemplo, imagine uma pesquisa de perfil socioeconômico que registre várias informações sobre chefes de família, tais como renda familiar em salários mínimos, posse de casa própria, posse de automóvel, posse de eletrodomésticos, entre outras. Imagine que um chefe de família tenha respondido o seguinte:

Renda	Casa própria	Número de automóveis	Viagem ao exterior	Quantos filhos?	Filhos estudam?
2 s.m.	Sim	3	2 vezes por ano	3	Escola particular

Isoladamente não há inconsistência ou erro, ou dado perdido, em cada uma das variáveis. Contudo ao comparar Renda às outras variáveis a existência de, no mínimo, um erro de registro é flagrante. Se, porém, não houve erro de registro (e o informante não for um megalômano mentiroso...), a renda realmente vale 2 salários mínimos, há uma inconsistência entre esta variável e todas as outras, pois não é possível[12] que alguém com tal renda consiga manter casa própria, 3 automóveis, 2 viagens por ano ao exterior, e 3 filhos estudando em escola particular. No exemplo acima, a inconsistência até que foi facilmente identificada, em outros, porém, são necessárias até mesmo técnicas avançadas de **mineração de dados** para descobri-las.

Para identificar inconsistências, especialmente aquelas derivadas de dados deliberadamente deturpados por um respondente, as pesquisas de opinião costumam incluir várias questões extras, que possibilitem cruzar respostas. No caso do exemplo acima, não se registra apenas a renda, mas outros aspectos que possibilitam caracterizar o padrão de vida do respondente, e, portanto, estimar qual é a sua renda real. A Receita Federal costuma utilizar procedimentos bastante sofisticados para identificar inconsistências, especialmente nas declarações de Imposto de Renda.

[11] Especialmente útil na detecção de fraudes, por exemplo, em telefonia celular um valor de conta muito acima do normal para certo usuário *pode* indicar a existência de "clonagem".
[12] A não ser que tenha perdido a sua principal fonte de renda recentemente, ou declare apenas a renda fixa.

2.2.5 – Recodificação

Algumas vezes é preciso criar novas variáveis a partir das existentes para facilitar a sua análise individual ou o cruzamento com outra para atingir os objetivos da análise. Isso pode ocorrer nos seguintes casos:
- uma variável qualitativa apresenta muitos valores possíveis, e há interesse em agrupá-los em um número menor de categorias (por exemplo, numa pesquisa de mobilidade urbana os diferentes bairros de uma mesma região geográfica podem ser agrupados);
- uma variável quantitativa será transformada em qualitativa (categorizada), através de uma expressão os valores quantitativos serão transformados em qualitativos (por exemplo, observações com idade entre 18 e 25 serão categorizadas como "jovens", idades entre 26 e 60 como "adultos" e acima de 60 como "idosos");
- uma variável quantitativa contínua será transformada em uma variável agrupada em classes (o procedimento será explicado na seção 2.2.3).

O processo de recodificação usualmente é feito em um aplicativo computacional (planilha eletrônica, gerenciador de banco de dados, sistema de informação gerencial ou software estatístico) utilizando alguma espécie de expressão lógica do tipo SE – ENTÃO – SENÃO, como pode ser mostrado na Figura 9, para o caso citado da idade.

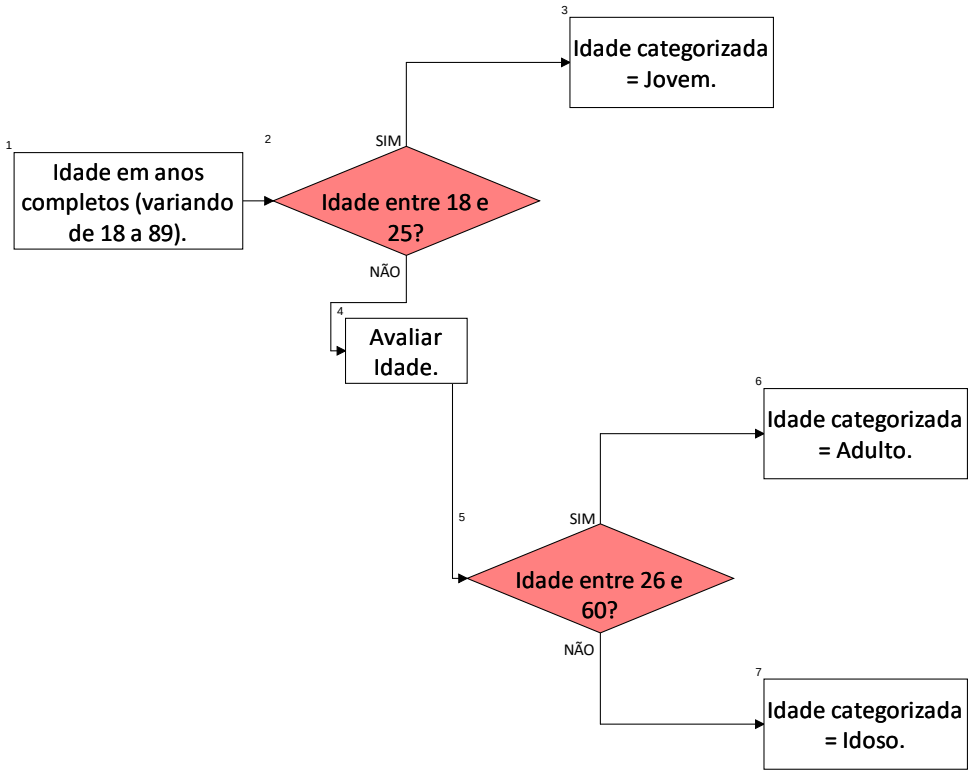


Figura 9 - Fluxograma de recodificação de uma variável quantitativa em uma qualitativa

Para cada valor de Idade é realizado o seguinte procedimento: primeiramente avalia-se se está entre 18 e 25 anos; caso esteja, o valor da variável qualitativa (categorizada) associado será "Jovem"; caso não esteja, prossegue-se com a avaliação; avalia-se então se a idade está entre 26 e 60 anos; caso esteja, o valor da variável qualitativa (categorizada) associado será "Adulto"; se a idade não se enquadrar nos dois critérios prévios, somente resta a opção de ser superior a 60 anos, então o valor da variável qualitativa (categorizada) associado será "Idoso".

O procedimento acima poderia ser modificado para um número maior ou menor de categorias, e poderia ser realizado com variáveis qualitativas também. Aqui cabe um aviso: a

categorização de uma variável quantitativa (transformação em variável qualitativa) é algo que deve ser evitado, pois uma variável quantitativa geralmente tem mais informação do que uma qualitativa. Mais adiante (seção 2.4) serão apresentadas as medidas de síntese que possibilitam descrever o comportamento de uma variável quantitativa de maneira bem objetiva, e um gráfico (seção 2.6) que permite uma visualização de vários aspectos de uma variável quantitativa, sem perda de informação.

2.2.6 – Transformação

A transformação é usualmente realizada apenas para variáveis quantitativas, quando há interesse em criar novas variáveis através de expressões matemáticas aplicadas à variável original. Isso é muito comum para operações financeiras (converter um valor monetário de uma moeda para outra – de dólar para real, real para dólar; converter de reais para salários mínimos ou vice-versa), conversão de unidades, cálculo de valores indexados, etc. Da mesma forma que na recodificação é recomendável o uso de aplicativos computacionais para tal procedimento.

Por exemplo, converter uma temperatura de Fahrenheit para Celsius pode ser feito pela seguinte equação:

$$Celsius = (Fahrenheit - 32) \times (5/9)$$

Esta equação pode ser implementada em planilha eletrônica permitindo obter uma nova variável temperatura em graus Celsius.

2.3 - Distribuição de frequências

O processo de resumo e organização dos dados busca basicamente registrar as ocorrências dos possíveis valores das variáveis que caracterizam o fenômeno, em suma consiste em elaborar DISTRIBUIÇÕES DE FREQUÊNCIAS das variáveis para que o conjunto de dados possa ser reduzido, possibilitando a sua análise.

A construção da distribuição de frequências exige que os possíveis valores da variável sejam discriminados e seja contado o número de vezes em que cada valor ocorreu no conjunto de dados. Para grandes arquivos de dados tal processo somente é viável utilizando meios computacionais.

Uma distribuição de frequências pode ser expressa em termos de tabelas ou através de gráficos, que terão algumas particularidades dependendo do nível de mensuração da variável.

2.3.1 - Distribuição de Frequências para variáveis qualitativas (nominais e ordinais)

a) Tabelas

Neste caso o número de possíveis realizações da variável costuma ser limitado, como na tabela do exemplo abaixo:

Exemplo. 2.2 - Usando os dados do Exemplo 2.1, empresa Escolástica Ltda., podemos construir tabelas de frequências para as variáveis sexo e função.
Supondo que haja 258 homens e 216 mulheres, 363 funcionários em escritório, 27 em serviços gerais, e 84 em gerência.

Tabela 1 - Sexo dos funcionários da empresa Escolástica Ltda.

Sexo	Frequência	Percentual
Masculino	258	54,43%
Feminino	216	45,57%
Total	474	100 %

Fonte: hipotética

Tabela 2 - Funções exercidas pelos funcionários da empresa Escolástica Ltda.

Função	Frequência	Percentual
Escritório	363	76,58%
Serviços gerais	27	5,70%
Gerência	84	17,72%
Total	474	100 %

Fonte: hipotética

As colunas Sexo e Função apresentam os possíveis valores que cada variável pode assumir, e a coluna frequência o número de ocorrências de cada um desses valores no conjunto de dados. Desta forma grandes conjuntos de dados podem ser ***resumidos*** *em pequenas tabelas. Usualmente calculam-se os percentuais de ocorrência de cada valor para permitir a COMPARAÇÃO COM CONJUNTOS DE DADOS DE TAMANHO DIFERENTE (onde a comparação direta das frequências pode levar a conclusões errôneas).*
O mais importante é interpretar as tabelas. Percebemos que não há grande diferença entre o percentual de homens e mulheres na empresa. Poderíamos concluir que não há predominância masculina significativa. Já na tabela das funções percebemos que a maioria esmagadora dos funcionários (76,58%) exerce atividades de escritório, restando 17,72% em gerência, e apenas 5,7% em serviços gerais (que talvez já estejam quase que totalmente terceirizados).

Obter as frequências de cada valor pode ser uma tarefa tediosa para grandes conjuntos de dados. Programas estatísticos ou mesmo planilhas eletrônicas permitem fazer tal contagem rapidamente e com menor chance de erro.

Um dos inconvenientes da utilização de tabelas para resumir conjuntos de dados é a demora na apreensão da informação: é necessário ler cada linha e coluna e posteriormente fazer o cruzamento das informações. Isso pode ser um problema em muitas situações em que há interesse em apresentar rapidamente as informações, então talvez a melhor forma de apresentar a distribuição de frequências seja através de um gráfico.

b) Gráficos

Dentre os vários gráficos disponíveis os mais utilizados para variáveis qualitativas são os gráficos de barras/colunas (bar chart) e os gráficos em setores (pie chart).

No gráfico de barras em um dos eixos são colocadas as categorias da variável e no outro as frequências ou percentuais de cada categoria. As barras podem ser horizontais ou verticais[13] (preferencialmente estas). Para os dados do Exemplo 2.2, usando as frequências, os gráficos seriam:

[13] Em português usualmente chama-se gráfico de barras quando estas são *horizontais*, e de gráfico de colunas quando elas são *verticais*, ao menos nas planilhas eletrônicas e softwares estatísticos.

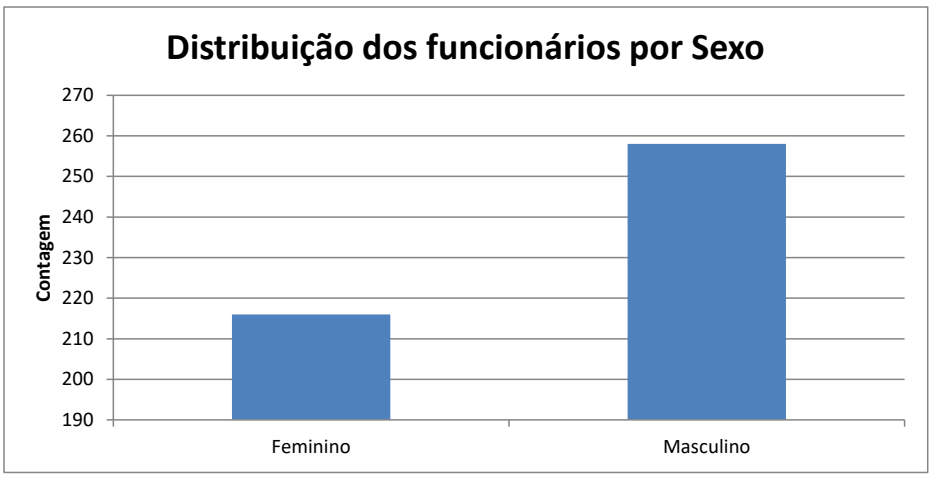


Figura 10 - Gráfico de colunas da variável Sexo (Escolástica Ltda.)

Fonte: hipotética

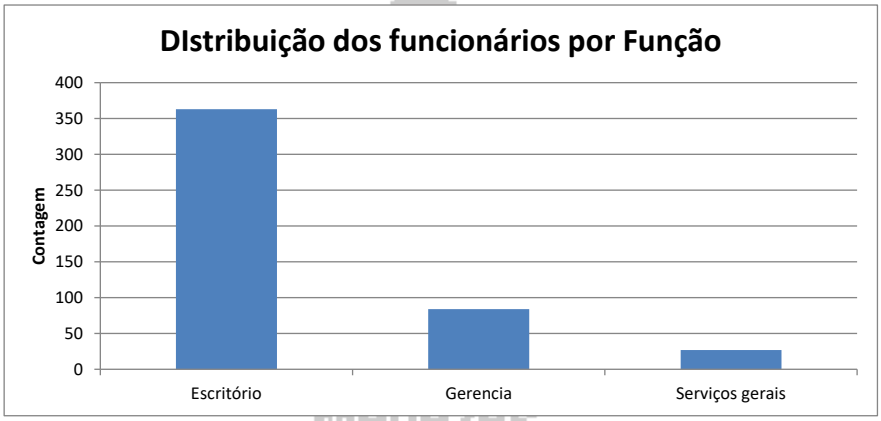


Figura 11 - Gráfico de barras de Função (Escolástica Ltda.)

Fonte: hipotética

Observe que a apreensão da informação da distribuição é bem mais fácil, percebe-se rapidamente na Figura 11 que há muito mais funcionários em Escritório do que nas outras funções. Contudo, na Figura 10 poderíamos ter a idéia de que há uma grande diferença entre os números de funcionários do sexo masculino e feminino: isso ocorre por causa da escala do gráfico, que começa em 190 (para evitar tal problema devemos, sempre que possível, iniciar a escala do gráfico de barras em zero). Este tipo de gráfico (para variáveis qualitativas) pode ser construído com a maioria dos processadores de texto e planilhas eletrônicas disponíveis atualmente.

O gráfico em setores, também chamado de gráfico circular ou em "pizza", exige uma construção mais sofisticada. Consiste em dividir um círculo (360^o) em setores proporcionais às realizações de cada categoria através de uma regra de três simples, na qual a frequência total (ou o percentual total 100%) corresponderia aos 360^o e a frequência ou a proporção de cada categoria corresponderia a um valor desconhecido em graus.

$$\text{Graus de uma categoria} = \frac{360^o \times \text{freq.(prop.) da categoria}}{\text{freq.(prop) total}}$$

Para os dados do Exemplo 2.1, para as variáveis sexo e função teríamos os seguintes valores:

Sexo

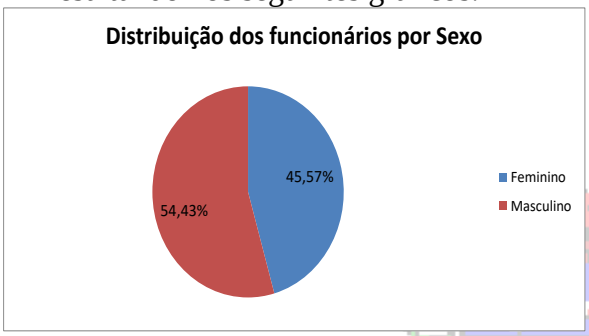
Masculino: $Graus = \frac{360° \times 258}{474} = 195,95°$ Feminino: $Graus = \frac{360° \times 216}{474} = 164,05°$

Função

Escritório: $Graus = \frac{360° \times 363}{474} = 275,7°$ Serviços gerais: $Graus = \frac{360° \times 27}{474} = 20,5°$

Gerência: $Graus = \frac{360° \times 84}{474} = 63,8°$

Resultando nos seguintes gráficos:



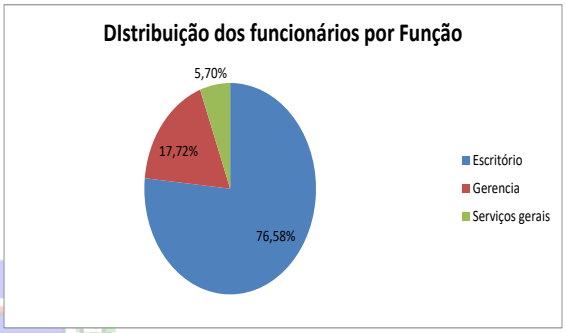


Figura 12 - Gráfico em setores de Sexo (Escolástica Ltda.) Fonte: hipotética

Figura 13 - Gráfico em setores de Função (Escolástica Ltda.) Fonte: hipotética

Tal como os gráficos de barras os gráficos de setores podem ser construídos por planilhas eletrônicas e mesmo processadores de texto.

c) Dupla classificação

Em todos os casos anteriores as distribuições de frequências referiam-se apenas a uma variável. Nas ciências econômicas e sociais é muito comum avaliar o comportamento conjunto de DUAS variáveis, através de uma dupla classificação. É muito comum representar essa distribuição conjunta de frequências através de uma **tabela de contingências**, para estudar a sua associação.

Exemplo 2.3 - Utilizando os dados do Exemplo 2.1, construir uma tabela de contingências para as variáveis Sexo e Função.

Seria necessário fazer o cruzamento das duas variáveis, anotando quantas ocorrências são verificadas em cada uma das combinações de valores possíveis: masculino - escritório, masculino - serviços gerais, masculino - gerência, feminino - escritório, feminino - serviços gerais, feminino-gerência. Dependendo do tamanho do conjunto de dados esta não é uma tarefa rápida. Vamos imaginar que obtivemos a tabela de contingência abaixo a partir de uma planilha eletrônica[14] (com os dados do Exemplo 2.1).

Tabela 3 - Tabulação cruzada de Sexo e Função (frequências) dos funcionários da empresa Escolástica Ltda.

Sexo	Função			
	Escritório	Serviços gerais	Gerência	Total
Masculino	157	27	74	258
Feminino	206	0	10	216
Total	363	27	84	474

Fonte: hipotética

Na função Escritório não há grande diferença (em termos relativos) entre homens e mulheres. Nas outras duas funções, todavia, o predomínio masculino é indiscutível, sendo especialmente importante nos cargos de gerência, onde as decisões são tomadas. A última coluna é o total marginal da variável Sexo (exatamente igual ao valor obtido no Exemplo 2.2), e a última linha é o

[14] Por exemplo, no Microsoft Excel é possível criar uma Tabela Dinâmica para fazer o cruzamento de variáveis.

total marginal da variável Função (tal como no Exemplo 2.2). Sem fazer o cruzamento entre as variáveis não conseguimos identificar o predomínio masculino, o que mostra a utilidade da tabela de contingências.

Os resultados poderiam ser apresentados em termos de percentuais, calculados em relação ao total geral, aos totais das linhas (totais dos valores de Sexo) ou aos totais das colunas (totais dos valores de Função), tal como mostrado a seguir:

Tabela 4 - Tabulação cruzada de Sexo e Função (% por coluna) dos funcionários da empresa Escolástica Ltda.

Sexo	Função			
	Escritório	Serviços gerais	Gerência	Total
Masculino	43,25%	100%	88,10%	54%
Feminino	56,75%	0%	11,90%	46%
Total	100%	100%	100%	100%

Fonte: hipotética

Observa-se que há apenas 11,90% de mulheres em cargos de gerência, havendo 46% de mulheres na organização (sem contar que há 0% de mulheres em serviços gerais), o que configura desigualdade de oportunidades, o que pode dar margem a diversas especulações.

Expressando a tabela em termos de percentuais por linha:

Tabela 5 - Tabulação cruzada de Sexo e Função (% por linha) dos funcionários da empresa Escolástica Ltda.

Sexo	Função			
	Escritório	Serviços gerais	Gerência	Total
Masculino	60,85%	10,47%	28,68%	100%
Feminino	95,37%	0,00%	4,63%	100%
Total	76,58%	5,70%	17,72%	100%

Fonte: hipotética

Observa-se que do total de mulheres apenas 4,63% exerce cargos de gerência, contra 28,68% dos homens.

A tabela de contingência poderia ser expressa em um gráfico composto de barras[15]*, sendo estes podendo ser de colunas agrupadas ou de colunas 100% empilhadas, ou por gráficos em setores apresentados conjuntamente*[16] *(dois gráficos de função, um para cada sexo, ou três gráficos de sexo, um para cada função). Os gráficos de colunas 100% empilhadas e os gráficos de setores apresentados conjuntamente permitem "ver" os percentuais por linha ou coluna mostrados na tabela de contingências.*

Vejam os gráficos a seguir:

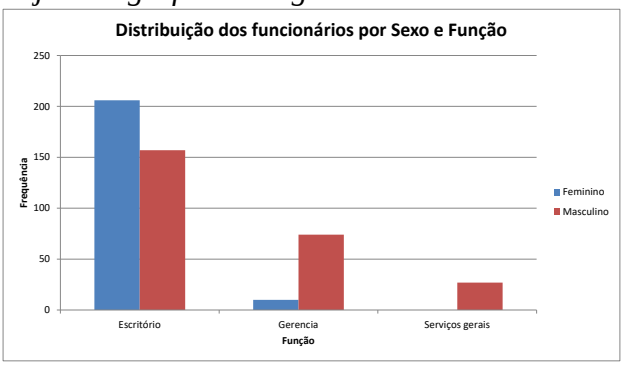


Figura 14 - Gráfico composto de colunas agrupadas

Fonte: hipotética

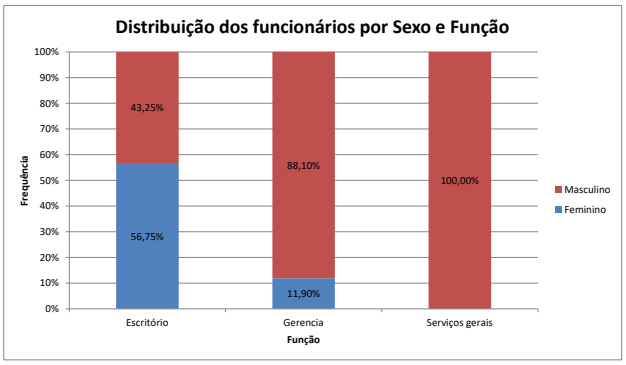


Figura 15 - Gráfico composto de colunas 100% empilhadas

Fonte: hipotética

[15] Pode ser construído em uma planilha eletrônica.
[16] Construídos por um programa estatístico.

A tabela de contingências com as frequências pode ser vista no gráfico da **Figura 14**. *A tabela com os percentuais por coluna pode ser vista no gráfico da* **Figura 15**.

No gráfico de colunas agrupadas as diferenças entre os grupos precisam ser bem grandes para que a visualização seja possível, e se houver um maior número de categorias (mais de quatro, por exemplo) em ambas as variáveis o gráfico ficará bastante congestionado.

No gráfico de colunas 100% empilhadas a escala sempre varia de zero a 100% (impossibilitando eventuais confusões) e quando as variáveis apresentam associação, no caso sexo e função, as colunas apresentam comportamento bem diferente, facilmente visualizável as diferenças de ocupação dos cargos por homens e mulheres em cada função. Mesmo nos casos em que há grande quantidade de categorias para cada variável a visualização permanece melhor do que no gráfico de colunas agrupadas. Na **Figura 16** *é possível observar o gráfico de colunas 100% empilhadas com os percentuais por linha: percebe-se claramente a diferença entre o percentual de homens e mulheres em cargos de gerência.*

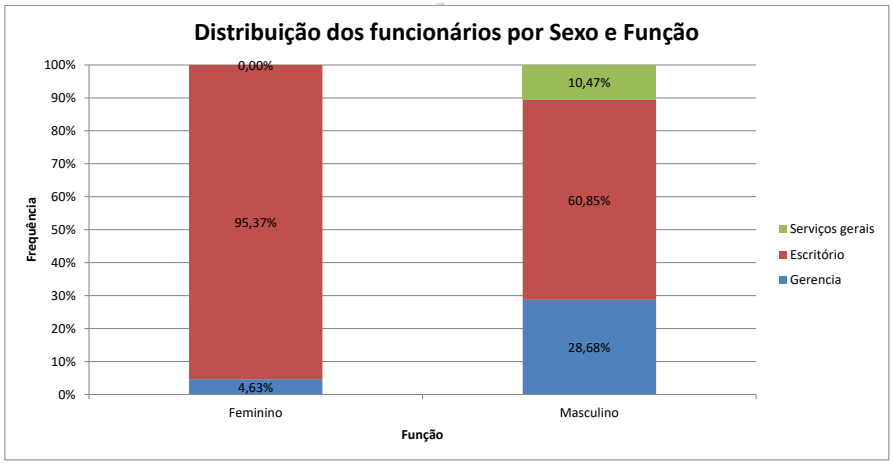


Figura 16 - Gráfico composto de colunas 100% empilhadas (% por linha)

Fonte: hipotética

2.3.2 - Distribuição de Frequências para variáveis quantitativas

A construção das distribuições de frequências para variáveis quantitativas é semelhante ao caso das variáveis qualitativas: relacionar os valores da variável com as suas ocorrências no conjunto de dados, mas apresenta alguns detalhes dependendo se a variável é discreta ou contínua.

Há várias formas de representar uma distribuição de frequências de variáveis quantitativas.

a) Rol ou lista

Consiste em ordenar os valores em ordem crescente ou decrescente, para que seja possível identificar valores extremos, mais comuns, etc. Somente é viável para pequenos conjuntos de dados.

b) Ramo-e-folhas (stem-and-leaf)

Trata-se de uma ferramenta exploratória útil para descrever pequenos conjuntos de dados (até 100 elementos). As observações são ordenadas crescentemente e "divididas" em duas partes para facilitar sua descrição:
Parte inteira | parte decimal ou Centena | dezena unidade decimais ou Milhar | centena (desprezando os demais algarismos) por exemplo.
Eventualmente alguns algarismos podem ser desprezados para facilitar a representação do conjunto. Uma das desvantagens do ramo e folhas é que ele não é único: dependendo do critério utilizado para definir os ramos e folhas a distribuição terá um aspecto diferente, o que pode modificar a

interpretação. Com o avanço da tecnologia computacional (especialmente na construção de gráficos) o ramo e folhas tornou-se menos importante.

Exemplo 2.4 - Construir o ramo e folhas das taxas de mortalidade infantil (por mil nascidos vivos) de alguns municípios do Oeste de SC:

32,3 62,2 10,3 22,0 13,1 9,9 11,9 20,0 36,4 23,5 18,0 22,6 20,3 38,3 19,6 27,2 28,9 18,4 27,3 21,7 23,7 13,9 36,3 32,9 29,7 25,4 23,8 15,7 17,0 39,2 22,7 29,9 18,3 33

Ramo e folhas das taxas mortalidade infantil no Oeste de SC.

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

Fonte: IBGE, GAPLAN- SC, 1987

*Para cada valor o primeiro algarismo é colocado à esquerda do traço vertical, no **ramo**, e o segundo algarismo à direita, nas **folhas**: o valor 32 passa a ser representado por 3 | 2. Observe também que as folhas estão ordenadas. Percebe-se a maior frequência de municípios com valores de taxa de mortalidade entre 10 e 20 por mil nascidos vivos (linhas 1 e 2), e claramente há um município discrepante (na linha 6) com taxa de 62,2 por mil nascidos vivos, o que deveria exigir uma investigação.*

c) Tabelas para dados não agrupados

Praticamente idênticas às das variáveis qualitativas, mas aqui as categorias são números. Basta contar quantas vezes cada valor ocorreu e registrá-lo (o valor original ou em percentual). Bastante utilizada para variáveis quantitativas DISCRETAS, para pequenos ou grandes conjuntos.

Exemplo 2.5 - Construir a tabela de frequências para dados não grupados para os valores a seguir:
Anos de educação dos funcionários da empresa Escolástica Ltda.

15	16	12	8	15	15	15	12	15	12	16	8	15	15	12	12	15	16	12	12	16
12	15	12	15	15	19	15	19	15	12	19	15	19	17	8	12	15	16	15	12	15
12	8	12	15	12	12	15	16	12	15	18	12	12	15	15	15	15	12	8	16	17
16	8	19	16	16	16	15	17	16	12	15	15	15	12	12	16	16	12	12	12	8
15	15	12	19	19	8	12	8	12	12	12	8	17	8	12	18	16	14	19	15	15
19	12	12	12	15	12	12	16	14	15	15	12	12	12	16	15	15	12	16	12	15
12	12	17	20	15	12	15	16	12	12	21	12	8	12	15	12	12	8	12	18	15
15	12	12	16	8	12	12	15	15	12	12	16	16	16	16	15	15	15	12	12	16
12	12	12	15	20	8	8	16	12	12	12	12	12	12	15	15	8	15	16	12	12
8	12	12	12	15	12	16	15	19	16	17	12	15	12	15	16	12	15	12	8	15
15	15	8	12	12	15	16	15	12	12	12	15	8	12	15	16	12	15	12	15	16
19	15	15	19	8	12	12	12	16	8	12	12	8	12	12	12	12	12	15	12	12
8	18	12	19	19	8	12	12	12	12	12	12	12	16	12	12	15	15	15	18	12
16	12	16	16	8	12	12	8	14	19	19	8	15	16	15	17	18	12	14	15	12
8	12	12	12	15	16	12	8	12	15	12	15	16	15	15	16	12	12	12	12	12
15	12	16	15	12	12	12	15	12	8	8	12	18	18	15	12	16	15	12	8	16
12	8	8	8	12	12	16	12	12	15	12	16	17	12	12	8	12	12	15	15	8
15	12	12	12	8	16	12	8	12	12	12	14	16	16	15	12	15	12	15	15	8
8	12	17	12	17	12	12	8	19	14	19	15	12	12	12	8	12	12	12	12	12
12	12	12	12	12	12	17	15	19	19	8	12	12	16	8	15	15	15	15	15	19
15	15	12	15	12	16	8	15	8	19	18	12	12	16	15	12	8	15	12	8	15
12	8	12	15	16	15	12	16	19	15	12	15	19	16	19	15	19	12	12	8	16
15	19	12	12	16	16	15	12	15	15	12	12	-	-	-	-	-	-	-	-	-

A variável quantitativa anos de educação pode assumir valores de 8 a 21: contou-se então o número de funcionários para cada valor possível. IMPORTANTE: para variáveis quantitativas

discretas devem ser incluídos todos os valores possíveis na tabela de frequências, incluindo os que têm frequência zero, no caso os valores 9, 10, 11 e 13. E podemos incluir na tabela a frequência e o percentual acumulados

Tabela 6 - Anos de educação dos funcionários da empresa Escolástica Ltda.

Anos de educação	Frequência	Frequência acumulada	%	% acumulado
8	53	53	11,18%	11,18%
9	0	53	0,00%	11,18%
10	0	53	0,00%	11,18%
11	0	53	0,00%	11,18%
12	190	243	40,08%	51,27%
13	0	243	0,00%	51,27%
14	6	249	1,27%	52,53%
15	116	365	24,47%	77,00%
16	59	424	12,45%	89,45%
17	11	435	2,32%	91,77%
18	9	444	1,90%	93,67%
19	27	471	5,70%	99,37%
20	2	473	0,42%	99,79%
21	1	474	0,21%	100,00%
Total	474	-	100%	-

Fonte: hipotética

O valor que ocorre com maior frequência é o 12, 40,08% dos funcionários da empresa tem 12 anos de educação (equivalente ao ensino médio completo). E, observando a coluna do % acumulado a maioria absoluta dos funcionários tem até 12 anos de educação: 51,27% tem no máximo 12 anos de educação. Então, deduz-se que 48,73% tem melhor qualificação, porque estudaram mais de 12 anos.

A tabela do Exemplo 2.5 poderia ser representada através de um **Histograma**, um gráfico de barras justapostas, em que as áreas das barras são proporcionais às frequências de cada valor.

Exemplo 2.6 - Representar a tabela de frequências do Exemplo 2.5 através de um histograma.

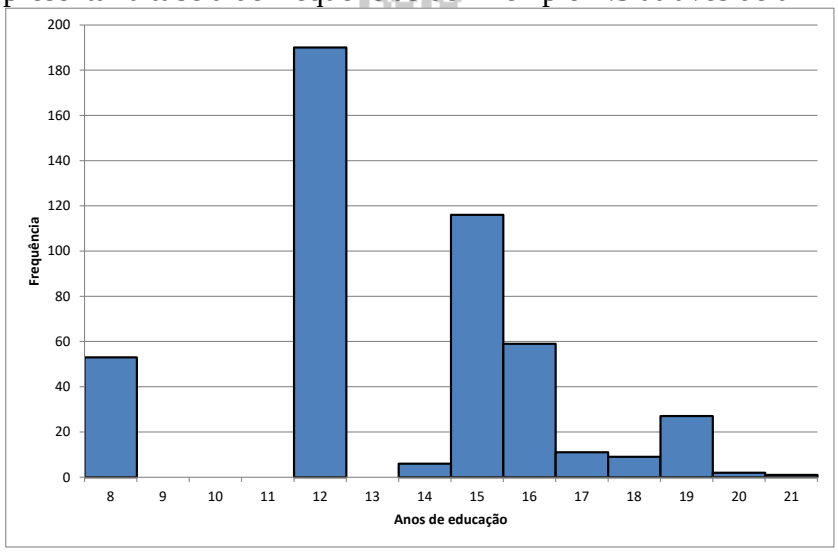


Figura 17 - Histograma para variável quantitativa discreta Anos de educação dos funcionários da empresa Escolástica Ltda.

Fonte: hipotética

d) Diagrama de pontos (dot-plot)

Recomendado para pequenos conjuntos de dados (até 100 elementos). Possibilita identificar valores discrepantes e avaliar a dispersão e do conjunto. Consiste em fazer com que cada resultado se identifique com um ponto na reta dos números reais: se o resultado repetir-se, acrescenta-se mais

um ponto no eixo vertical do gráfico, tantas vezes quantas o resultado ocorrer. É uma ferramenta mais apropriada para variáveis quantitativas CONTÍNUAS (onde os valores ocorrem apenas uma vez ou poucas vezes). Uma das vantagens do diagrama de pontos é que ele é único para um conjunto de dados.

Exemplo 2.7 - Construir o diagrama de pontos para as taxas percentuais de crescimento demográfico de alguns municípios catarinenses:
-0,4 -1,7 -1,0 0,3 -0,3 -0,45 -0,15 -1,2 -0,1 -0,42 0,6 0,4 7,3 3,6 -0,6 3,2 6,6 3,0 2,9 2,4

É preciso ordenar os dados do menor para o maior, e registrar as ocorrências de cada um na reta dos reais. Observe que se trata de uma variável quantitativa CONTÍNUA: crescimento demográfico expresso em números, e pode assumir uma infinidade de valores (negativos, significando redução da população, zero, significando estagnação, positivos, aumento da população). Nenhum valor ocorre mais de uma vez. O gráfico resultante pode ser visto na Figura 18.

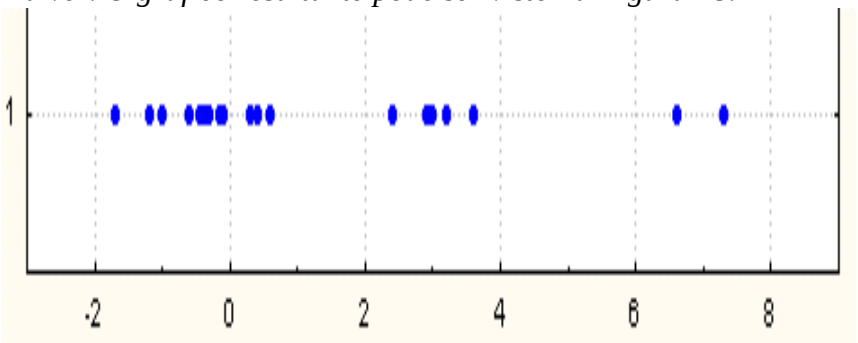


Figura 18 - Diagrama de pontos para variável quantitativa contínua taxa de crescimento demográfico de municípios catarinenses

Fonte: hipotética

Analisando o gráfico da Figura 18 podem-se identificar claramente três grupos de municípios. O primeiro é composto por municípios com crescimento negativo ou pouco acima de zero, que constituem a maioria. Por que isso ocorre? Identificando os municípios, e estudando suas características socioeconômicas poder-se-ia responder. O segundo grupo é formado por municípios com crescimento "intermediário", entre 2 e 4%. E o terceiro grupo é formado por valores discrepantes, com crescimento acima de 6% (o que pode acarretar graves problemas de infraestrutura). Todas essas informações foram obtidas de um gráfico extremamente simples. Não obstante poucos programas estatísticos apresentam o diagrama de pontos como opção.

e) Tabelas para dados agrupados

Quando as variáveis são contínuas sua grande variação torna inúteis as tabelas para dados não agrupados como instrumento de resumo do conjunto, pois praticamente todos os valores têm frequência baixa, o que resultaria em uma tabela enorme. É preciso representar os dados através de um conjunto de classes mutuamente exclusivas (para que cada valor pertença apenas a uma classe), que contenha do menor ao maior valor do conjunto: registram-se então quantos valores do conjunto estão em cada classe. É mais apropriada para grandes conjuntos de dados.

O processo para montagem das classes é o seguinte:

1) Determinar o intervalo do conjunto (diferença entre o maior e o menor valor do conjunto).

2) Dividir o intervalo em um número conveniente de classes k, onde:

- para até 100 observações $k = \sqrt{No\ de\ elementos}$;
- para mais de 100 observações $k = 5 \times log(n)$

3) Obter a amplitude de cada classe c: $c = Intervalo/k$

Muito provavelmente c será um valor fracionário, mas o arredondamento deve ser feito *para cima*, para garantir que todos os valores do conjunto serão considerados na tabela.

4) Estabelecer as classes com a seguinte notação:

Li - limite inferior Ls - limite superior Li |-- Ls limite inferior incluído, superior excluído

Li |--| Ls ambos incluídos (geralmente usado apenas na última classe, quando o limite superior coincide com o maior valor do conjunto.

5) Determinar as frequências de cada classe.

6) Determinar os pontos médios de cada classe através da média dos 2 limites (serão os representantes das classes).

O processo acima é mostrado na Figura 19:

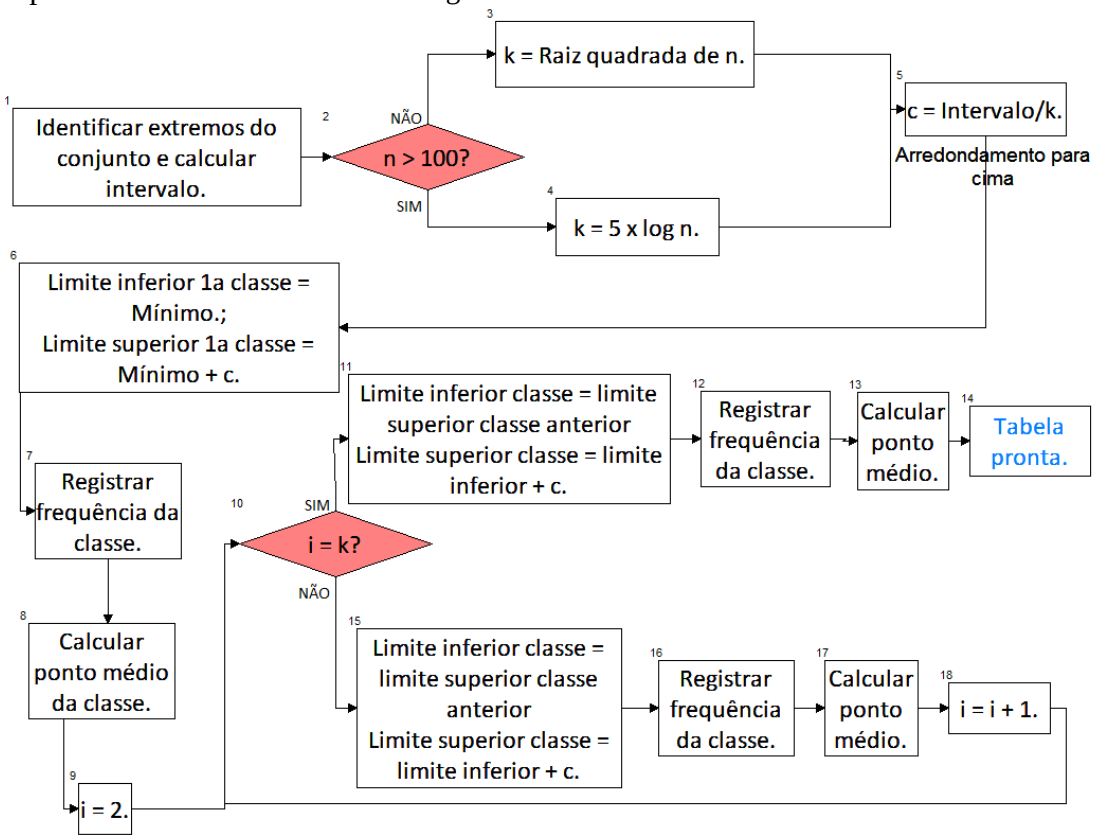


Figura 19 - Processo de construção de uma tabela de dados agrupados

Exemplo 2.8 - Construir a tabela de frequências agrupada em classes para os dados do salário anual (em US$) dos funcionários da Escolástica Ltda. mostrada abaixo:

15750	15900	16200	16200	16200	16350	16500	16650	16800	16950
16950	16950	17100	17100	17250	17400	17400	17700	18150	18150
18450	18750	19200	19200	19650	19650	19650	19650	19650	19650
19800	19950	19950	20100	20100	20400	20400	20400	20550	20700
20700	20850	20850	20850	20850	20850	21000	21000	21150	21150
21300	21300	21300	21300	21450	21450	21450	21600	21600	21600
21750	21750	21750	21900	21900	21900	21900	21900	22050	22050
22050	22050	22200	22200	22200	22350	22350	22350	22350	22350
22350	22500	22500	22500	22500	22500	22500	22500	22650	22650
22800	22800	22950	22950	22950	22950	23100	23100	23100	23100
23250	23250	23400	23400	23400	23400	23550	23550	23700	23700
23850	23850	24000	24000	24000	24000	24000	24000	24000	24000
24150	24150	24150	24150	24300	24300	24300	24300	24450	24450
24450	24450	24450	24450	24450	24450	24600	24600	24750	24750
24750	24750	24900	25050	25050	25050	25050	25200	25200	25200
25200	25200	25350	25350	25500	25500	25500	25500	25650	25650
25800	25800	25950	25950	25950	25950	26100	26100	26250	26250
26250	26250	26250	26250	26250	26400	26400	26400	26400	26550
26550	26550	26550	26550	26700	26700	26700	26700	26700	26700

26700	26850	26850	27000	27000	27000	27150	27150	27300	27300
27300	27300	27300	27450	27450	27450	27450	27450	27600	27600
27750	27750	27750	27750	27750	27750	27750	27900	27900	27900
27900	28050	28050	28050	28200	28350	28350	28350	28500	28500
28500	28500	28500	28500	28650	28800	28800	28950	29100	29100
29100	29100	29100	29160	29250	29250	29340	29400	29400	29400
29400	29400	29550	29700	29850	29850	29850	29850	30000	30000
30000	30000	30000	30150	30150	30270	30300	30300	30300	30300
30450	30600	30600	30600	30750	30750	30750	30750	30750	30750
30750	30750	30750	30750	30750	30750	30750	30900	30900	30900
31050	31200	31200	31200	31350	31350	31350	31500	31500	31500
31650	31650	31650	31650	31950	31950	31950	31950	32100	32400
32550	32550	32550	32850	33000	33150	33300	33300	33300	33450
33540	33750	33900	33900	33900	33900	33900	33900	34410	34500
34500	34500	34500	34500	34620	34800	34800	34950	35100	35100
35250	35250	35250	35250	35550	35550	35700	35700	35700	36000
36000	36000	36000	36150	36600	37050	37500	37650	37800	37800
37800	38400	38550	38700	38850	38850	39150	39300	39600	39900
40050	40200	40200	40200	40200	40350	40350	40800	40800	41100
41550	42000	42300	42300	43000	43410	43500	43650	43950	44875
45000	45150	45250	45625	46000	46000	46875	47250	47550	48000
48750	49000	50000	50550	51000	51250	51450	52125	52650	53125
54000	54375	54875	54900	55000	55000	55000	55500	55750	56500
56550	56750	57000	58125	58750	59375	59400	60000	60000	60375
60625	61250	61875	61875	62500	65000	65000	65000	66000	66250
66750	66875	66875	67500	68125	68125	68750	68750	69250	70000
70000	70875	72500	73500	73750	75000	75000	78125	78250	78500
80000	81250	82500	83750	86250	90625	91250	92000	97000	100000
103500	103750	110625	135000	-	-	-	-	-	-

Resolução:

1) *Intervalo= Maior - Menor = 135000 - 15750 = 119250 (o maior salário foi de 135000 e o menor de 15750, as classes devem englobar do menor ao maior valor).*

2) *Como há 474 elementos no conjunto:* $No\ de\ classes = 5 \times log(n) = 5 \times log(474) = 13{,}379 \cong 13$ *(usando as regras de arredondamento vistas anteriormente obtém-se o número de classes).*

3) *Amplitude das classes = 119250/13 = 9173,077*

(como há um intervalo de 119250 e 13 classes, e a variável não tem casas decimais, a amplitude arredondada será de 9174; é importante que neste caso o arredondamento seja feito SEMPRE PARA CIMA, para ***<u>garantir</u>*** *que todos os valores sejam considerados na tabela).*

4) *Classes:*

15750|--24924
24924|--34098
34098|--43272
43272|--52446
52446|--61620
61620|--70794
70794|--79968
79968|--89142
89142|--98316
98316|--107490
107490|--116664
116664|--125838
125838|--135012

5) *Pontos médios de cada classe: (limite inferior + limite superior)/2*

6) *Frequências de cada classe (é recomendável no mínimo o uso de uma planilha eletrônica, tabela dinâmica, para grandes quantidades de dados): podemos incluir percentuais, frequência acumulada e percentuais acumulados.*

Tabela 7 - Salário anual (em US$) dos funcionários da Escolástica Ltda.

Classes	*Pontos médios*	*Frequência*	*%*	*Frequência acumulada*	*% acumulado*
15750|--24924	*20337*	*143*	*30,17%*	*143*	*30,17%*
24924|--34098	*29511*	*185*	*39,03%*	*328*	*69,20%*
34098|--43272	*38685*	*57*	*12,03%*	*385*	*81,22%*
43272|--52446	*47859*	*23*	*4,85%*	*408*	*86,08%*
52446|--61620	*57033*	*24*	*5,06%*	*432*	*91,14%*
61620|--70794	*66207*	*19*	*4,01%*	*451*	*95,15%*
70794|--79968	*75381*	*9*	*1,90%*	*460*	*97,05%*
79968|--89142	*84555*	*5*	*1,05%*	*465*	*98,10%*
89142|--98316	*93729*	*4*	*0,84%*	*469*	*98,95%*
98316|--107490	*102903*	*3*	*0,63%*	*472*	*99,58%*
107490|--116664	*112077*	*1*	*0,21%*	*473*	*99,79%*
116664|--125838	*121251*	*0*	*0,00%*	*473*	*99,79%*
125838|--135012	*130425*	*1*	*0,21%*	*474*	*100,00%*
Total		*474*	*100%*	-	-

Fonte: hipotética

Observe que a classe de 24924 a 34098 dólares apresenta a maior frequência, 185, correspondendo a 39,03% do total. Se olharmos para a coluna do % acumulado, constata-se que 69,20% dos funcionários da Escolástica Ltda. recebem um salário anual de no máximo 34098 dólares. Há apenas um funcionário com salário entre107490 e 116664 dólares, e também apenas um na última classe (125838 a 135012 dólares, o valor máximo 135000 dólares). PERDEU-SE informação sobre o conjunto original: sabe-se que há 9 salários anuais entre 70794 e 79968 dólares, mas não se mais quais são os seus valores exatos, ou seja, as frequências das classes passam a ser as frequências dos pontos médios.

Os pontos médios nem sempre são os representantes mais fiéis das classes. Para uma grande quantidade de dados há maior probabilidade de que estas estimativas representem os verdadeiros valores, mas as medidas estatísticas calculadas com base na tabela de frequências agrupada em classes serão apenas *estimativas* dos valores reais devido à perda de informação referida acima.

A tabela do Exemplo 2.8 também pode ser representada através de um histograma. O número de barras é igual ao número de classes. Cada barra é centrada no ponto médio de cada classe, e o ponto inicial de cada barra é o limite inferior da classe, e o ponto final é o limite superior.

Exemplo 2.9 - Representar a tabela de frequências do Exemplo 2.8 através de um histograma.

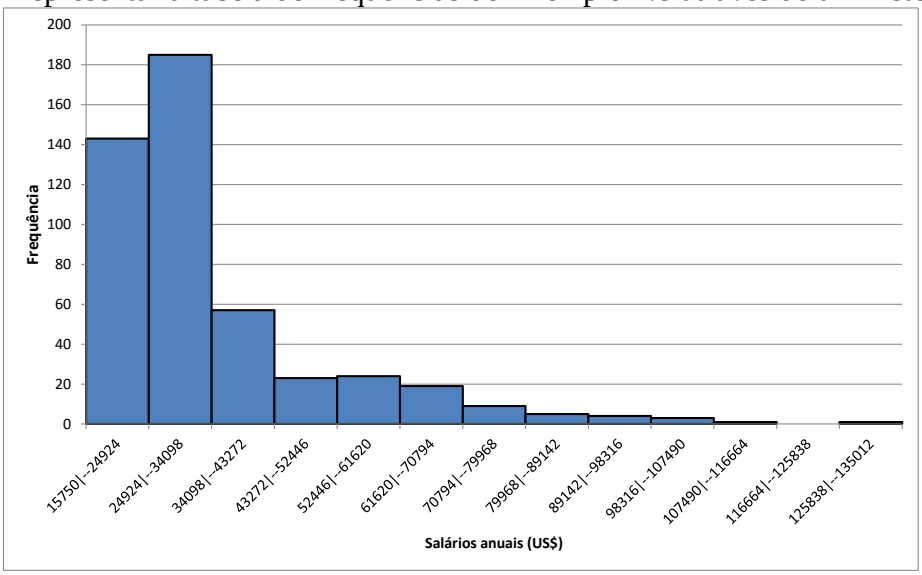


Figura 20 - Histograma para variável quantitativa contínua Salário anual (em US$) dos funcionários da Escolástica Ltda.

Fonte: hipotética

A apreensão da informação é muito mais rápida, percebe-se que as três primeiras classes (até 43272 dólares) concentram a maioria dos funcionários. Este comportamento, com maior concentração de frequências nos menores valores é típico de variáveis como renda, salário, gastos, enfim, variáveis envolvendo quantidades monetárias. É importante observar que se um número diferente de classes fosse escolhido arbitrariamente (maior ou menor) o aspecto da tabela agrupada e do histograma agrupado poderiam ficar bem diferentes, o que pode ser um inconveniente na interpretação.

2.4 - Medidas de Síntese (Estatísticas)

Vimos anteriormente que um conjunto de dados pode ser resumido através de uma distribuição de frequências, e que esta pode ser representada através de uma tabela ou de um gráfico. Se o conjunto refere-se a uma variável QUANTITATIVA (Intervalar) há uma terceira maneira de resumi-lo: as Medidas de Síntese. A vantagem das medidas de síntese é a sua objetividade, para a maioria delas há apenas um valor possível para um conjunto de dados específico, o que torna a interpretação dos seus resultados mais direta.

As Medidas de Síntese, também chamadas de Estatísticas, dividem-se em Medidas de Posição (Medidas de Tendência Central), Medidas de Dispersão e Separatrizes.

As Medidas de Posição obtém um valor numérico que represente a tendência do conjunto (valor "típico"). As mais importantes são: Média, Mediana, e Moda.

As Medidas de Dispersão obtém uma mensuração da disposição dos dados no conjunto, da sua variabilidade (se estão concentrados em torno de um valor, se distribuídos, etc). As mais importantes são: Intervalo, Variância, Desvio Padrão e Coeficiente de Variação.

As Separatrizes dão medidas que dividem o conjunto em um certo número de partes iguais: Quartis (4 partes), Decis (10 partes), Centis (100 partes).

Vamos ver cada uma em profundidade.

2.4.1 - Medidas de Posição

As Medidas de Posição procuram caracterizar a tendência central do conjunto, um valor numérico que "represente" o conjunto. Esse valor pode ser calculado levando em conta todos os valores do conjunto ou apenas alguns valores ordenados.

Média ($\bar{X}$)

A Média aqui citada é a média aritmética simples, a soma dos valores observados dividida pelo número desses valores. Seja um conjunto de **n** valores de uma variável quantitativa X, a média do conjunto será:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$ Onde x_i é um valor qualquer do conjunto e $\sum_{i=1}^{n} x_i$ é a soma dos valores do conjunto.

Exemplo 2.10 - A tabela abaixo se refere às notas finais de três turmas de estudantes. Calcular a média de cada turma:

Turma	Valores
A	4 5 5 6 6 7 7 8
B	1 2 4 6 6 9 10 10
C	0 6 6 7 7 7 7,5 7,5

Ao somar os valores em cada turma teremos o mesmo resultado: 48. Como cada turma tem 8 alunos as três turmas terão a mesma média: 6.

No exemplo acima as três turmas têm a mesma média (6), então se apenas essa medida fosse utilizada para caracterizá-las poderíamos ter a impressão que as três turmas têm desempenhos idênticos. Será? Observe atentamente a tabela acima.

Na primeira turma temos realmente os dados distribuídos regularmente em torno da média, com a mesma variação tanto abaixo quanto acima. Já na segunda vemos uma distorção maior, embora a maioria das notas sejam altas algumas notas baixas "puxam" a média para um valor menor. E no terceiro grupo há apenas uma nota baixa, mas seu valor é tal que realmente consegue diminuir a média do conjunto.

Um dos problemas da utilização da média é que, por levar em conta TODOS os valores do conjunto, ela pode ser distorcida por valores discrepantes ("outliers") que nele existam. É importante então interpretar corretamente o valor da média.

O valor da média pode ser visto como o ponto central de cada conjunto de dados, ou seja, o ponto de equilíbrio do conjunto: "se os valores do conjunto fossem pesos sobre uma tábua, a média é a posição em que um suporte equilibra esta tábua".

Vamos ver como os valores do exemplo distribuem-se em um diagrama apropriado:

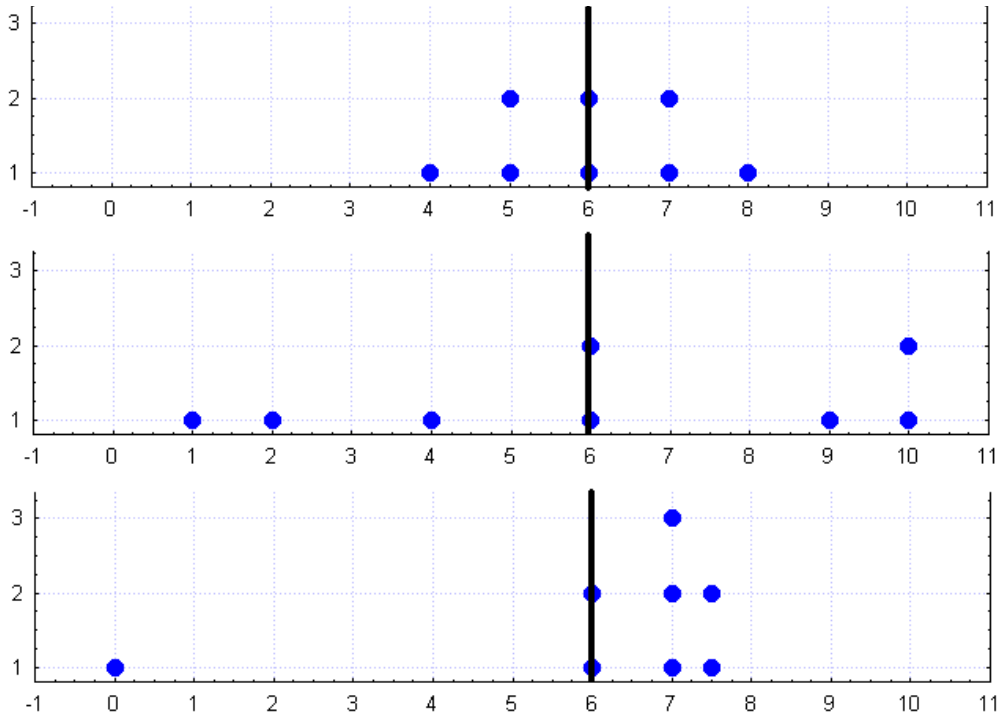


Figura 21 - Interpretação do valor da média

A média dos três conjuntos é a mesma, mas observe as diferentes disposições dos dados. O primeiro grupo apresenta os dados distribuídos de forma simétrica em torno da média. No segundo grupo a distribuição já é mais irregular, com valores mais "distantes" na parte de baixo, e o terceiro grupo é claramente assimétrico em relação à média (que foi distorcida pelo valor discrepante 0). Portanto muito cuidado ao caracterizar um conjunto apenas por sua média[17].

Outro aspecto importante a ressaltar é que a média pode ser um valor que a variável ***não pode assumir***. Isto é especialmente verdade para variáveis quantitativas discretas, resultantes de contagem, como número de filhos, quando a média pode assumir um valor "quebrado", 4,3 filhos, por exemplo.

Já foi muito comum[18] calcular médias de variáveis quantitativas a partir de distribuições de frequências representadas em tabelas: simplesmente multiplica-se cada valor (ou o ponto médio da classe) pela frequência associada, somam-se os resultados e divide-se o somatório pelo número de observações do conjunto. Na realidade trata-se de uma média **ponderada** pelas frequências de ocorrência de cada valor da variável.

$$\overline{x} = \frac{\sum_{i=1}^{k} (x_i \times f_i)}{n}$$

Onde k é o número de valores da variável discreta, ou o número de classes da variável agrupada, e x_i é um valor qualquer da variável discreta, ou o ponto médio de uma classe qualquer.

Exemplo 2.11 - Calcular a média dos anos de educação para os dados do Exemplo 2.5

Tabela 8 - Anos de educação dos funcionários da empresa Escolástica Ltda.

Anos de educação (x_i)	Frequência (f_i)	$x_i \times f_i$
8	53	424
9	0	0
10	0	0
11	0	0
12	190	2280
13	0	0
14	6	84
15	116	1740
16	59	944
17	11	187
18	9	162
19	27	513
20	2	40
21	1	21
Total	474	6395

Fonte: hipotética

Neste caso a média do conjunto será: $\overline{x} = \frac{\sum_{i=1}^{14} (x_i \times f_i)}{n} = \frac{6395}{474} \cong 13{,}49$ *anos de educação.*

Observe que há 14 valores (k =14) diferentes na tabela do Exemplo 2.5. NENHUMA pessoa pode ter 13,49 anos de serviço naquele conjunto de dados. Assim, não se esqueça de que a média pode assumir valores que a variável não pode assumir. O valor 13,49 está situado entre os valores com

[17] Essa era a grande crítica que era feita nas décadas de 60 e 70 sobre as medições de nível de desenvolvimento. Era comum medir o nível de desenvolvimento de um país por sua renda per capita (PIB/número de habitantes), uma média, que não revelava, porém, a CONCENTRAÇÃO de renda do país, levando a conclusões errôneas sobre a qualidade de vida em muitos países.

[18] Ainda é, em algumas provas de concursos públicos.

maiores frequências: 12 (com frequência igual a 190) e 15 (com frequência igual a 116), o valor 13 tem frequência zero, e 13,49 é mais próximo de 12 do que de 15 (por ter maior frequência o valor 12 "puxa" a média para ele), mas os valores superiores de anos de educação (15, 16, 19) acabam também afastando a média do valor 12 que tem a maior frequência. Na Figura 22 *é possível visualizar a posição da média, como centro de massa.*

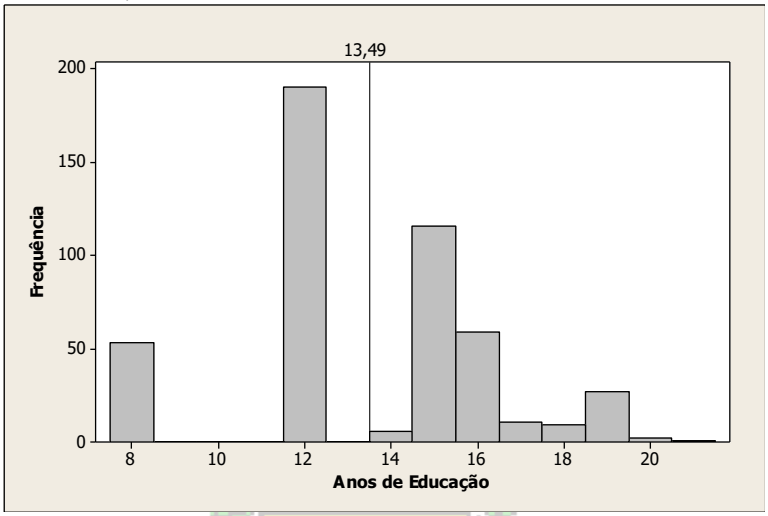


Figura 22 – Posição na média no histograma de anos de educação

Exemplo 2.12- Calcular a média dos salários anuais dos funcionários da empresa Escolástica Ltda., com base na tabela do Exemplo 2.8

Tabela 9 - Salário anual (em US$) dos funcionários da Escolástica Ltda.

Classes	*Pontos médios – x_i*	*Frequência - f_i*	$x_i \times f_i$
15750|--24924	*20337*	*143*	*2908191*
24924|--34098	*29511*	*185*	*5459535*
34098|--43272	*38685*	*57*	*2205045*
43272|--52446	*47859*	*23*	*1100757*
52446|--61620	*57033*	*24*	*1368792*
61620|--70794	*66207*	*19*	*1257933*
70794|--79968	*75381*	*9*	*678429*
79968|--89142	*84555*	*5*	*422775*
89142|--98316	*93729*	*4*	*374916*
98316|--107490	*102903*	*3*	*308709*
107490|--116664	*112077*	*1*	*112077*
116664|--125838	*121251*	*0*	*0*
125838|--135012	*130425*	*1*	*130425*
Total		*474*	*16327584*

Fonte: hipotética

Observe que há 13 classes (k = 13).

Neste caso a média do conjunto será: $\bar{x} = \dfrac{\sum_{i=1}^{13}(x_i \times f_i)}{n} = \dfrac{16327584}{474} \cong 34446{,}38$ *dólares*

Repare que a média está na classe de 34098 a 43272 dólares, a terceira classe, com frequência igual a 57, e não nas duas primeiras classes que têm frequências bem maiores. Isso ocorre porque os valores maiores de salário anual, que podem chegar a 130425 dólares (ponto médio da última classe) distorcem a média para cima (vimos no Exemplo 2.8 que 69,20% dos funcionários da Escolástica Ltda. recebem um salário anual de no máximo 34098 dólares, e constata-se neste exemplo que a média está acima disso).

Quando os dados não estão grupados (Exemplo 2.11) o resultado será idêntico ao que seria obtido simplesmente somando todos os valores e dividindo o somatório pelo número de valores. Contudo, se a tabela estiver agrupada em classes (Exemplo 2.12) TODAS as medidas (não somente a média) serão apenas estimativas dos valores reais, pois as medidas serão calculadas usando os pontos médios (que são os representantes das classes) e não mais os valores originais. No caso do Exemplo 2.12 a média real vale 34419,57 dólares.

Atualmente com as facilidades computacionais disponíveis não se calcula mais a média (ou qualquer outra medida) a partir de uma tabela agrupada em classes se os dados originais estão disponíveis: os programas calculam as medidas usando os dados originais e as tabelas são apresentadas apenas para dar uma idéia da variação dos dados. Na planilha eletrônica Microsoft Excel ® em português calcula-se a média com a seguinte função: MÉDIA(intervalo com as células[19]).

NÃO CALCULE NENHUMA MEDIDA ESTATÍSTICA COM BASE EM UMA TABELA AGRUPADA EM CLASSES SE VOCÊ TIVER ACESSO AOS DADOS ORIGINAIS!

Mediana (Md)

A mediana é o ponto que divide o conjunto em duas partes iguais: metade dos dados têm valor menor do que a mediana e a outra metade têm valor maior do que a mediana[20]. Pouco afetada por eventuais ***valores discrepantes*** existentes no conjunto (que costumam distorcer substancialmente o valor da média).

"A mediana de um conjunto de valores é o valor que ocupa a posição (**n** +1)/2, quando os dados estão ***ordenados*** crescente ou decrescentemente.

Se (**n** +1)/2 for fracionário toma-se como mediana a média dos dois valores que estão nas posições imediatamente abaixo e acima de (**n** +1)/2".

Exemplo 2.13 - Calcular a mediana para as notas das três turmas do Exemplo 2.10.

Turma	Valores
A	4 5 5 6 6 7 7 8
B	1 2 4 6 6 9 10 10
C	0 6 6 7 7 7 7,5 7,5

*Posição mediana = (**n** + 1)/2 = (8+1)/2 = 4,5ª significa que o valor da mediana será calculado através da média entre os valores que estiverem na 4ª e na 5ª posição do conjunto[21].*

Turma A: Md = (6 + 6)/ 2 = 6
Turma B: Md = (6 + 6)/ 2 = 6
Turma C: Md = (7 + 7)/ 2 = 7

Observe que a mediana da Turma C é diferente, mais alta, refletindo melhor o conjunto de dados, uma vez que há apenas uma nota baixa. Perceba também que apenas os dois valores centrais foram considerados para obter a mediana, deixando o resultado "imune" aos valores discrepantes.

[19] O intervalo com as células registra a localização (coluna e linha) onde estão as células na planilha com os valores para os quais se pretende calcular a média, ou qualquer outra medida. As planilhas eletrônicas usualmente identificam as células com letras nas colunas e números nas linhas.
[20] Então a mediana também é uma SEPARATRIZ.
[21] Por esse motivo os dados PRECISAM estar ordenados crescentemente.

Exemplo 2.14 - Calcular a mediana para o grupo a seguir:

10 11 12 13 15 16 16 35 60

*Posição mediana = (**n** + 1)/2 = (9+1)/2 = 5ª como o conjunto tem um número ímpar de valores o valor da mediana será igual ao valor que estiver na 5ª posição.*

Md = 15 Média = 20,89

Observe que neste caso média e mediana são diferentes, pois a média foi distorcida pelos valores mais altos 35 e 60, que constituem uma minoria. Neste caso a medida de posição que melhor representaria o conjunto seria a mediana. Se a média é diferente da mediana a distribuição da variável quantitativa no conjunto de dados é dita ASSIMÉTRICA[22].

Tal como a média a mediana pode ser calculada a partir de uma tabela de frequências, com as mesmas ressalvas feitas para aquela medida. Os programas estatísticos, e muitas planilhas eletrônicas dispõem de funções que calculam a mediana. Na planilha eletrônica Microsoft Excel ® em português calcula-se a mediana com a seguinte função: MED(intervalo com as células).

Exemplo 2.15 - Calcule a mediana do salário anual dos funcionários da empresa Escolástica Ltda., de acordo com a tabela do Exemplo 2.8.

Tabela 10 - Salário anual (em US$) dos funcionários da Escolástica Ltda.

Classes	*Pontos médios*	*Frequência*	*%*	*Frequência acumulada*
15750|--24924	*20337*	*143*	*30,17%*	*143*
24924|--34098	*29511*	*185*	*39,03%*	*328*
34098|--43272	*38685*	*57*	*12,03%*	*385*
43272|--52446	*47859*	*23*	*4,85%*	*408*
52446|--61620	*57033*	*24*	*5,06%*	*432*
61620|--70794	*66207*	*19*	*4,01%*	*451*
70794|--79968	*75381*	*9*	*1,90%*	*460*
79968|--89142	*84555*	*5*	*1,05%*	*465*
89142|--98316	*93729*	*4*	*0,84%*	*469*
98316|--107490	*102903*	*3*	*0,63%*	*472*
107490|--116664	*112077*	*1*	*0,21%*	*473*
116664|--125838	*121251*	*0*	*0,00%*	*473*
125838|--135012	*130425*	*1*	*0,21%*	*474*
Total		*474*	*100%*	-

Fonte: hipotética

O primeiro passo é encontrar a posição da mediana:

*Posição mediana = (**n** + 1)/2 = (474+1)/2 = 237,5ª*

Como o resultado não é inteiro, precisamos fazer a média entre os valores que estão nas posições imediatamente anterior e posterior à mediana: 237ª e 238ª respectivamente. Como a tabela está agrupada em classes é como se os pontos médios fossem os valores dos conjuntos.

Na coluna "frequência acumulada" podemos identificar quais os valores que estão na 237ª e 238ª posições. Observe que na primeira classe há 143 ocorrências: o ponto médio 20337 dólares ocorre 143 vezes (ocupando então as posições 1ª a 143ª). A frequência acumulada até a segunda classe vale 328 (ou seja, o último valor da segunda classe ocupa a 328ª posição).

Como estamos procurando os valores que ocupam as 2377ª e 238ª posições, e a primeira classe vai até a 143ª posição, e a segunda classe vai até a 328ª posição (e começa na 144ª), os valores que ocupam as 237ª e 238ª são iguais a 29511 dólares, o ponto médio da segunda classe.

Então a mediana será:

Md = (29511 + 29511)/ 2 = 29511 dólares

Novamente o valor acima é apenas uma estimativa, a mediana real vale:

Md = (28800 + 28950)/2 = 28875 dólares

[22] Maiores detalhes serão apresentados no Diagrama em Caixas.

Moda (Mo)

A moda é o valor da variável que ocorre com maior frequência no conjunto.

É a medida de posição de obtenção mais simples, e também pode ser usada para variáveis qualitativas, pois apenas registra qual é o valor mais freqüente, podendo este valor ser tanto um número quanto uma categoria de uma variável nominal ou ordinal.

Um conjunto pode ter apenas uma Moda, várias Modas ou nenhuma Moda.

Exemplo 2.16 - Encontre a moda das notas das três turmas do Exemplo 2.10.

Turma	Valores
A	4 5 5 6 6 7 7 8
B	1 2 4 6 6 9 10 10
C	0 6 6 7 7 7 7,5 7,5

A turma A tem 3 modas: os valores 5, 6 e 7 ocorrem duas vezes cada. A turma B tem duas modas: os valores 6 e 10 ocorrem duas vezes cada. A turma C tem uma moda apenas: o valor 7 ocorre 3 vezes.

Podemos apresentar uma breve comparação das medidas de posição.

Propriedades	Média	Mediana	Moda
Definição	$\overline{x} = \frac{\sum x}{n}$	Valor do meio	Valor mais freqüente
Existência	Sempre existe	Sempre existe	Pode não existir, pode haver mais de uma
Leva em conta todos os valores	Sim	Não	Não
Afetada por valores discrepantes	Sim	Não	Não
Vantagens	Usada em muitos métodos estatísticos	Menos sensível a valores discrepantes	Apropriada para dados qualitativos

2.4.2 - Medidas de Dispersão

O objetivo das medidas de dispersão é medir quão próximos uns dos outros estão os valores de um grupo (e algumas mensuram a dispersão dos dados em torno de uma medida de posição).

Intervalo

É a medida mais simples de dispersão. Consiste em identificar os valores extremos do conjunto (mínimo e máximo), podendo ser expresso:

- pela diferença entre o valor máximo e o mínimo;
- pela simples identificação dos valores.

Exemplo 2.17 - Obtenha o Intervalo para os conjuntos de notas das duas turmas abaixo:

Turma	Valores
A	4 5 5 6 6 7 7 8
B	4 4 4,2 4,3 4,5 5 5 8

O intervalo será o mesmo para ambas as turmas: [4,8] ou 4.

Observe que no Exemplo 2.17 as duas turmas apresentam o mesmo intervalo (4). Mas observando os dados percebe-se facilmente que a dispersão dos dados tem comportamento diferente nas duas turmas, e essa á principal desvantagem do uso do intervalo como medida de dispersão.

Se colocarmos os dados do Exemplo 2.17 em um diagrama apropriado:

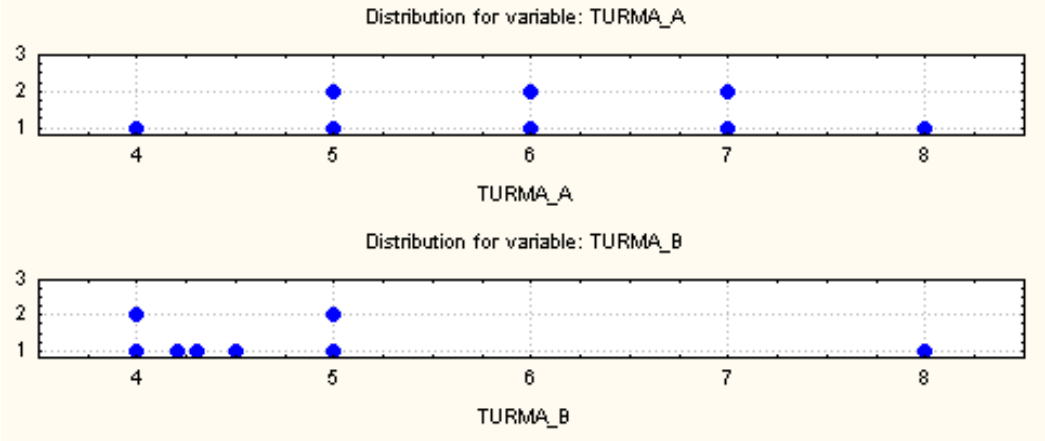


Figura 23 - Desvantagem do uso do intervalo como medida de dispersão

Observa-se na Figura 23 que os dados da turma A apresentam uma dispersão bem mais uniforme do que os da turma B, embora ambos os conjuntos tenham o mesmo intervalo. O intervalo não permite ter idéia de como os dados estão distribuídos ENTRE os extremos (não permite identificar que o valor 8 na turma B é um valor discrepante).

Na planilha eletrônica Microsoft Excel ® em português pode-se identificar o intervalo obtendo os valores máximo e mínimo do conjunto com as seguintes funções:

MÁXIMO(intervalo com as células) MÍNIMO(intervalo com as células)

Variância (s^2)

A variância é uma das medidas de dispersão mais importantes. É a média aritmética dos quadrados dos desvios de cada valor em relação à **média**[23]: proporciona uma mensuração da dispersão dos dados em torno da média.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad \text{(amostra)}$$

Onde x_i é um valor qualquer do conjunto. Se os dados referem-se a uma POPULAÇÃO usa-se **n** no denominador da expressão[24].

A unidade da variância é o quadrado da unidade dos dados (e, portanto, o quadrado da unidade da média) causando dificuldades para avaliar a dispersão: se, por exemplo, a variável peso com média de 75 kg em um conjunto de dados, e ao calcular a variância obtemos 12 kg^2 a avaliação da dispersão torna-se difícil. Não obstante, a variância e a média são as medidas geralmente usadas para caracterizar as distribuições probabilísticas (que serão vistas adiante).

[23] Então, *conceitualmente*, é preciso calcular a média *antes* de obter a variância.

[24] A razão dessa distinção será explicada no item Inferência Estatística. Pode-se adiantar que a utilização de **n - 1** no denominador é *indispensável* para que a variância da variável na amostra possa ser um bom estimador da variância da variável na população. A maioria dos programas computacionais, porém, costuma calcular o desvio padrão supondo que os dados são provenientes de uma amostra. Em algumas planilhas eletrônicas há funções pré-programadas para ambos os casos.

O que se pode afirmar, porém, é que quanto maior a variância, mais dispersos os dados estão em torno da média (maior a dispersão do conjunto).

Para fins de Análise Exploratória de Dados, caracterizar a dispersão através da variância não é muito adequado. Costuma-se usar-se a raiz quadrada positiva da variância, o desvio padrão.

Desvio Padrão (s)

É a raiz quadrada positiva da variância, apresentando a mesma unidade dos dados e da média, permitindo avaliar melhor a dispersão.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad \text{(amostra)}$$

As mesmas observações sobre população e amostra feitas para a variância são válidas para o desvio padrão. Nas calculadoras que dispõe de modo estatístico é comum que os botões com as fórmulas de desvio padrão sejam da seguinte forma:

- s_x ou σ_{n-1} para desvio padrão amostral (com n-1 no denominador);
- σ_x ou σ_n para desvio padrão populacional (com n no denominador)

É prática comum ao resumir através de medidas de síntese um conjunto de dados referente a uma variável quantitativa apresentar a média e o desvio padrão desse conjunto, para que seja possível ter uma idéia do valor típico e da distribuição dos dados em torno dele.

O desvio padrão pode assumir valores menores do que a média, da mesma ordem de grandeza da média, ou até mesmo maiores do que a média.

A fórmula anterior costuma levar a consideráveis erros de arredondamento basicamente porque usa o valor da média. Se o valor desta for uma dízima, um arredondamento terá que ser feito, causando um pequeno erro, e este erro será propagado pelas várias operações de subtração (de cada valor em relação à média) e potenciação (elevação ao quadrado da diferença entre cada valor e a média). A fórmula modificada a seguir reduz o erro de arredondamento:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i^2) - \left[\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}{n-1}} \quad \textit{(amostra)}$$

Primeiramente cada valor (x_i) do conjunto é elevado ao quadrado e somam-se todos os resultados obtendo $\sum_{i=1}^{n} x_i^2$. Somam-se também todos os valores do conjunto para obter $\sum_{i=1}^{n} x_i$, somatório este que será elevado ao quadrado. Os somatórios e o valor de **n** (número de elementos no conjunto) são substituídos na fórmula para obter os resultados[25].

Exemplo 2.18 – Calcule o desvio padrão para as notas das três turmas do Exemplo 2.10, supondo que sejam amostras.

Turma	Valores
A	4 5 5 6 6 7 7 8
B	1 2 4 6 6 9 10 10
C	0 6 6 7 7 7 7,5 7,5

[25] É desta forma que os programas computacionais calculam o desvio padrão.

Podemos obter os desvios padrões usando as duas fórmulas (a original e a modificada). Lembrando que as três turmas tem a MESMA média ($\bar{x} = 6$).

Turma	*Valores*									*Somas*
A	x_i	4	5	5	6	6	7	7	8	48
	x_i^2	16	25	25	36	36	49	49	64	300
	$x_i - \bar{x}$	-2	-1	-1	0	0	1	1	2	0
	$(x_i - \bar{x})^2$	4	1	1	0	0	1	1	4	12
B	x_i	1	2	4	6	6	9	10	10	48
	x_i^2	1	4	16	36	36	81	100	100	374
	$x_i - \bar{x}$	-5	-4	-2	0	0	3	4	4	0
	$(x_i - \bar{x})^2$	25	16	4	0	0	9	16	16	86
C	x_i	0	6	6	7	7	7	7,5	7,5	48
	x_i^2	0	36	36	49	49	49	56,25	56,25	331,5
	$x_i - \bar{x}$	-6	0	0	1	1	1	1,5	1,5	0
	$(x_i - \bar{x})^2$	36	0	0	1	1	1	2,25	2,25	43,5

Observe que a soma dos desvios em relação à média ($x_i - \bar{x}$) é SEMPRE igual a zero, porque a média é o centro de massa do conjunto de dados, por isso, para obter uma medida mensurável de dispersão os desvios precisam ser elevados ao quadrado[26].

Substituindo as somas nas fórmulas dos desvios padrões para as três turmas:

Turma	*Fórmula conceitual*	*Fórmula modificada*
A	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{12}{8-1}} \cong 1{,}31$	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i^2) - \left[\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}{n-1}} = \sqrt{\frac{300 - \left[\frac{48^2}{8}\right]}{8-1}} \cong 1{,}31$
B	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{86}{8-1}} \cong 3{,}51$	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i^2) - \left[\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}{n-1}} = \sqrt{\frac{374 - \left[\frac{48^2}{8}\right]}{8-1}} \cong 3{,}51$
C	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{43{,}5}{8-1}} \cong 2{,}49$	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i^2) - \left[\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}{n-1}} = \sqrt{\frac{331{,}5 - \left[\frac{48^2}{8}\right]}{8-1}} \cong 2{,}49$

Observe que os resultados dos desvios padrões por ambas as fórmulas são iguais porque apenas duas notas na turma C são fracionárias, todas as outras são números naturais. Como as médias das notas das turmas são IGUAIS é possível comparar os desvios padrões diretamente:

[26] Ou ter seu sinal removido, passando a ser desvio absoluto, mas o desvio quadrático é mais usado porque a variância e o desvio padrão têm muito mais importância na Estatística.

- a turma A tem o menor desvio padrão (1,31), pois suas notas tem a menor dispersão, variam de 4 a 8 (intervalo igual a 4);
- a turma B tem o maior desvio padrão (3,51), pois suas notas tem a maior dispersão, variam de 1 a 10 (intervalo igual a 9);
- a turma C tem um desvio padrão intermediário (2,49), pois suas notas variam entre 0 e 7,5 (intervalo igual a 7,5) – se não fosse pela nota zero, teria a menor dispersão.[27]

Tal como no caso da média pode haver interesse em calcular o desvio padrão de variáveis quantitativas a partir de distribuições de frequências representadas em tabelas. Os valores da variável (ou os pontos médios das classes), e os quadrados desses valores serão multiplicados por suas respectivas frequências:

$$s = \sqrt{\frac{\sum_{i=1}^{k}\left(x_i^{\,2} \times f_i\right) - \left[\frac{\left(\sum_{i=1}^{k} x_i \times f_i\right)^2}{n}\right]}{n-1}} \quad \text{(amostra)}$$

Onde $\mathbf{x_i}$ é o valor da variável ou ponto médio da classe, $\mathbf{f_i}$ a frequência associada, e k é o número de valores da variável discreta, ou o número de classes da variável agrupada.

Exemplo 2.19 - Calcular o desvio padrão do salário anual dos funcionários da empresa Escolástica Ltda. a partir da tabela do Exemplo 2.8 – supor que é uma população.

Tabela 11 - Salário anual (em US$) dos funcionários da Escolástica Ltda.

Classes	*Pontos médios – x_i*	*x_i^2*	*Frequência - f_i*	*$x_i \times f_i$*	*$x_i^2 \times f_i$*
15750|--24924	*20337*	*413593569*	*143*	*2908191*	*59143880367*
24924|--34098	*29511*	*870899121*	*185*	*5459535*	*161116337385*
34098|--43272	*38685*	*1496529225*	*57*	*2205045*	*85302165825*
43272|--52446	*47859*	*2290483881*	*23*	*1100757*	*52681129263*
52446|--61620	*57033*	*3252763089*	*24*	*1368792*	*78066314136*
61620|--70794	*66207*	*4383366849*	*19*	*1257933*	*83283970131*
70794|--79968	*75381*	*5682295161*	*9*	*678429*	*51140656449*
79968|--89142	*84555*	*7149548025*	*5*	*422775*	*35747740125*
89142|--98316	*93729*	*8785125441*	*4*	*374916*	*35140501764*
98316|--107490	*102903*	*10589027409*	*3*	*308709*	*31767082227*
107490|--116664	*112077*	*12561253929*	*1*	*112077*	*12561253929*
116664|--125838	*121251*	*14701805001*	*0*	*0*	*0*
125838|--135012	*130425*	*17010680625*	*1*	*130425*	*17010680625*
Total		-	*474*	*16327584*	*702961712226*

Fonte: hipotética

$$\sum_{i=1}^{13}\left(x_i^{\,2} \times f_i\right) = 7029617122\ 26 \qquad \sum_{i=1}^{13} x_i \times f_i = 16327584$$

Onde k = 13 (há 13 classes) e n = 474 (há 474 observações)

$$s = \sqrt{\frac{\sum_{i=1}^{k}\left(x_i^{\,2} \times f_i\right) - \left[\frac{\left(\sum_{i=1}^{k} x_i \times f_i\right)^2}{n}\right]}{n}} = \sqrt{\frac{702961712226 - \left[\frac{(16327584)^2}{474}\right]}{474}} \cong 17218,84$$

[27] Seu desvio padrão seria igual a 0,63.

Tal como na média, o resultado do desvio padrão calculado através de uma tabela agrupada em classes será apenas uma estimativa do valor real (que foi igual a 25106,02 dólares): SEMPRE QUE HOUVER ACESSO AOS DADOS ORIGINAIS (DADOS BRUTOS) AS MEDIDAS DE SÍNTESE DEVEM SER CALCULADAS A PARTIR DESTES.

Na planilha eletrônica Microsoft Excel ® em português calcula-se o desvio padrão com as seguintes funções:
Desvio padrão amostral DESVPAD(intervalo com as células)
Desvio padrão populacional DESVPADP(intervalo com as células)

Coeficiente de Variação Percentual (c.v.%)

O coeficiente de variação percentual é uma medida de dispersão relativa, pois permite comparar a dispersão de diferentes distribuições (com diferentes médias e desvios padrões).

$$c.v.\% = \frac{s}{\bar{x}} \times 100\%$$

Onde **s** é o desvio padrão da variável no conjunto de dados, e $\bar{x}$ é a média no mesmo conjunto.

Quanto menor o coeficiente de variação percentual, mais os dados estão concentrados em torno da média, pois o desvio padrão é pequeno em relação à média. Esta medida não funciona muito bem quando a média é negativa ou muito próxima de zero: uma média negativa levaria a um coeficiente de variação negativo, e muito próxima de zero resultaria em um CV% "artificialmente" muito grande.

Exemplo 2.20 Calcular o coeficiente de variação percentual para as notas das turmas do Exemplo 2.10, e indicar qual das três apresenta as notas mais homogêneas.

Turma	Valores
A	4 5 5 6 6 7 7 8
B	1 2 4 6 6 9 10 10
C	0 6 6 7 7 7 7,5 7,5

Para a turma A: $\bar{x} = 6$ $s = 1,31$ $c.v.\% = (1,31/6) \times 100 = 21,82\%$

Para a turma B: $\bar{x} = 6$ $s = 3,51$ $c.v.\% = (3,51/6) \times 100 = 58,42\%$

Para a turma C: $\bar{x} = 6$ $s = 2,49$ $c.v.\% = (2,49/6) \times 100 = 41,55\%$

A turma mais homogênea é a A, pois apresenta o menor cv% das três. Isso era esperado, uma vez que as notas da turma A estão distribuídas mais regularmente do que as das outras. A turma B tem a maior dispersão por apresentar valores mais distantes da média (1 e 10) e a turma C tem dispersão intermediária por causa da nota zero.

No caso acima a comparação ficou ainda mais simples, pois as médias dos grupos eram iguais, bastaria avaliar os desvios padrões dos grupos, mas para comparar a dispersão de distribuições com médias diferentes deve-se usar o coeficiente de variação percentual.

Teorema (Desigualdade) de Chebyshev

Por que é tão importante calcular a média e o desvio padrão dos valores de uma variável em um conjunto de dados? Há um teorema que permite, a partir da média e do desvio padrão, obter estimativas dos extremos do conjunto (muito útil para os casos de amostra). Formalmente:

"A proporção (ou fração) de ***qualquer*** conjunto de dados a menos de K desvios padrões a contar da média é sempre ***ao menos*** $1 - 1/K^2$, onde K é um número positivo maior do que 1."

Traduzindo:

- para K = 2, pelo teorema de Chebyshev, $1 - 1/K^2 = 0,75$; então ao menos 3/4 (75%) de todos os elementos do conjunto estão no intervalo que vai de 2 desvios padrões abaixo da média a 2 desvios padrões acima da média.
- para K = 3, pelo teorema de Chebyshev, $1 - 1/K^2 = 0,89$; então ao menos 8/9 (89%) de todos os

elementos do conjunto estão no intervalo que vai de 3 desvios padrões abaixo da média a 3 desvios padrões acima da média.

Exemplo 2.21 - Uma pesquisa por amostragem identificou que a renda mensal de um estado apresenta média de 1800 reais e desvio padrão de 200 reais. Usando o teorema de Chebyshev identifique os limites estimados onde estão pelo menos 75% das rendas.
Conforme visto anteriormente, se a proporção de interesse é 0,75 (75%), então K será igual a 2. Assim, podemos encontrar os valores que estão a 2 desvios padrões da média:
- 2 desvios padrões abaixo = 1800 - 2 × 200 = 1400 reais
- 2 desvios padrões acima = 800 + 2 × 200 = 2200 reais.
Então pelo menos 75% das rendas mensais devem estar entre 400 e 1200 reais.

Na prática as proporções reais costumam ser maiores do que os valores calculados pelo Teorema de Chebyshev. Mas, o Teorema apresenta a vantagem de ser válido para todos os casos, e não exigir o conhecimento da distribuição seguida pelos dados para estimar as proporções, basta apenas o cálculo da média e do desvio padrão.

2.4.3 - Separatrizes

As separatrizes são valores que dividem a distribuição em um certo número de partes iguais: a **mediana** divide em 2 partes iguais, os **quartis** dividem em 4 partes iguais, os decis em 10 partes iguais e os centis em 100 partes iguais. O objetivo é proporcionar uma melhor idéia da dispersão do conjunto, principalmente da *simetria* ou *assimetria* da distribuição. Vamos nos limitar aos quartis.

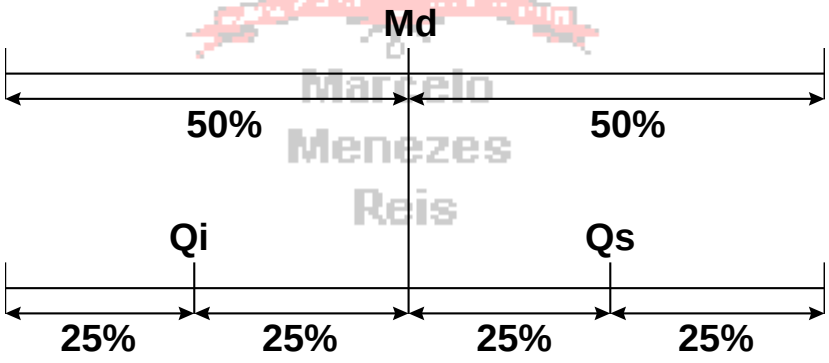


Figura 24 - Percentuais de dados definidos pelas separatrizes principais

Quartis

Os quartis são as separatrizes que dividem o conjunto em 4 partes iguais.

O primeiro quartil ou quartil inferior (**Qi**) é o valor do conjunto que delimita os 25% menores valores: 25% dos valores são menores do que **Qi** e 75% são maiores do que **Qi**.

O segundo quartil ou quartil do meio é a própria mediana (**Md**), que separa os 50% menores dos 50% maiores valores.

O terceiro quartil ou quartil superior (**Qs**) é o valor que delimita os 25% maiores valores: 75% dos valores são menores do que **Qs** e 25% são maiores do que **Qs**.[28]

Como são medidas baseadas na ordenação dos dados, o conjunto precisa ser previamente ordenado, e é necessário que primeiramente sejam calculadas as posições dos quartis.

Posição do quartil inferior = **(n + 1)/4** Posição do quartil superior = **[3×(n+1)]/4**

Após calcular a posição encontrar o elemento do conjunto que nela está localizado, o que dependerá de o número de observações for ímpar ou par:

[28] O quartil inferior corresponde ao 25º centil, a mediana corresponde ao 50º centil e o quartil superior corresponde ao 75º centil.

- se as posições forem valores *inteiros*, basta recuperar os valores que estão nas posições;
- se as posições forem valores *fracionários*, identificar os valores na posição imediatamente anterior (val_ant) e posterior (val_post)[29],

1º caso

a posição do quartil tem parte fracionária igual a ,25, para obter o valor do quartil
=> Quartil inferior/superior = val_ant × 0,75 + val_post × 0,25

2º caso

a posição do quartil tem parte fracionária igual a ,5, para obter o valor do quartil
=> Quartil inferior/superior = val_ant × 0,5 + val_post × 0,5

3º caso

a posição do quartil tem parte fracionária igual a ,75, para obter o valor do quartil
=> Quartil inferior/superior = val_ant × 0,25 + val_post × 0,75

Se os dados estiverem dispostos em uma distribuição de frequências, utilizar o mesmo procedimento observando as frequências associadas a cada valor (variável discreta) ou ponto médio de classe.

Exemplo 2.22 – Calcular mediana, quartil inferior e quartil superior do gasto total anual (em US$) de 99 famílias (dados de 1980) mostradas na tabela a seguir.

No.	Gasto	No.	Gasto	No.	Gasto	No.	Gasto	No.	Gasto	No.	Gasto	No.	Gasto
1	4.328	16	8.060	31	12.184	46	15.557	61	20.398	76	28.860	91	43.043
2	4.393	17	8.634	32	12.185	47	15.900	62	20.441	77	29.747	92	44.000
3	4.428	18	8.688	33	12.411	48	16.059	63	21.070	78	31.715	93	44.321
4	4.825	19	8.827	34	12.542	49	16.241	64	21.723	79	32.865	94	44.884
5	4.858	20	9.006	35	12.828	50	16.864	65	21.866	80	33.330	95	47.859
6	4.928	21	9.213	36	13.105	51	17.304	66	21.918	81	33.798	96	47.868
7	5.537	22	9.370	37	13.200	52	17.876	67	22.338	82	34.671	97	48.816
8	5.660	23	9.742	38	13.252	53	18.110	68	24.029	83	34.857	98	50.718
9	5.870	24	9.903	39	13.632	54	18.114	69	24.509	84	34.984	99	53.413
10	6.286	25	10.451	40	13.941	55	18.664	70	25.158	85	35.797	-	-
11	6.352	26	10.454	41	14.012	56	18.838	71	25.432	86	36.667	-	-
12	7.428	27	11.304	42	14.660	57	18.895	72	25.606	87	37.111	-	-
13	7.468	28	11.430	43	14.858	58	18.977	73	25.895	88	39.263	-	-
14	7.619	29	11.560	44	14.909	59	19.349	74	26.267	89	40.878	-	-
15	7.766	30	11.707	45	15.010	60	19.663	75	26.644	90	41.189	-	-

Como há 99 famílias, n = 99. Podemos calcular as posições da mediana, quartil inferior e quartil superior:
Pos Md = (n + 1)/2 = (99 + 1)/2 = 50ª. Pos Qi = (n + 1)/4 = (99 + 1)/4 = 25ª.
Pos Qs = 3 × (n + 1)/4 = 3 × (99 + 1)/4 = 75ª.
Como os valores são todos inteiros e a tabela acima está ordenada é possível recuperar diretamente os valores: Md = 16.864 dólares Qi = 10.451 dólares Qs = 26.644 dólares
Então 50% das famílias tinham gastos anuais de até 16.864 dólares e 50% acima deste valor. E 25% das famílias tinham gastos anuais de até 10.451 dólares e 75% acima deste valor. E ainda 75% das famílias tinham gastos anuais de até 26.644 dólares e 25% acima deste valor.

Exemplo 2.23 – Repetir o Exemplo 2.22 supondo que fosse acrescentada mais uma família, a 100ª com gasto anual de 54.125 dólares.
Como agora há 100 famílias, n = 100. Podemos calcular as posições da mediana, quartil inferior e quartil superior:
Pos Md = (n + 1)/2 = (100 + 1)/2 = 50,5ª. Pos Qi = (n + 1)/4 = (100 + 1)/4 = 25,25ª.
Pos Qs = 3 × (n + 1)/4 = 3 × (100 + 1)/4 = 75,75ª.

[29] Este é um dos métodos usados para cálculo dos quartis para conjuntos com número *par* de observações, procedimento também usado na função QUARTIL.EXC do Microsoft Excel ®.

Mediana: como a posição vale 50,5 é preciso calcular a média entre os valores que estão nas posições 50ª e 51ª .

Valor na posição 50ª. = 16.864 valor na posição 51ª. = 17.304

Md = (16.864+17.304)/2 = 17.084 dólares

Quartil inferior: como a posição vale 25,25 é preciso identificar os valores que estão nas posições 25ª e 26ª.

Valor na posição 25ª = 10.451 valor na posição 26ª = 10.454

Qi = 10.451 × 0,75 + 10.454 × 0,25 = 10.451,75 dólares

Quartil superior: como a posição vale 75,75 é preciso identificar os valores que estão nas posições 75ª e 76ª.

Valor na posição 75ª = 26.644 valor na posição 76ª = 28.860

Qs = 26.644 × 0,25 + 28.860 × 0,75 = 28.306 dólares

Exemplo 2.24 - Calcular o quartil inferior e o quartil superior para o salário anual dos funcionários da empresa Escolástica Ltda. a partir da tabela do Exemplo 2.8.

Tabela 12 - Salário anual (em US$) dos funcionários da Escolástica Ltda.

Classes	*Pontos médios*	*Frequência*	*%*	*Frequência acumulada*
15750|--24924	*20337*	*143*	*30,17%*	*143*
24924|--34098	*29511*	*185*	*39,03%*	*328*
34098|--43272	*38685*	*57*	*12,03%*	*385*
43272|--52446	*47859*	*23*	*4,85%*	*408*
52446|--61620	*57033*	*24*	*5,06%*	*432*
61620|--70794	*66207*	*19*	*4,01%*	*451*
70794|--79968	*75381*	*9*	*1,90%*	*460*
79968|--89142	*84555*	*5*	*1,05%*	*465*
89142|--98316	*93729*	*4*	*0,84%*	*469*
98316|--107490	*102903*	*3*	*0,63%*	*472*
107490|--116664	*112077*	*1*	*0,21%*	*473*
116664|--125838	*121251*	*0*	*0,00%*	*473*
125838|--135012	*130425*	*1*	*0,21%*	*474*
Total		*474*	*100%*	-

Fonte: hipotética

Por estar em uma tabela de frequências o conjunto já está ordenado. Calculando as posições dos quartis:

Pos Qi = (n + 1)/4 = (474 + 1)/4 = 118,75ª.

Pos Qs = 3 × (n + 1)/4 = 3 × (474 + 1)/4 = 356,25ª.

Quartil inferior: como a posição vale 118,75ª é preciso identificar os valores que estão nas posições 118ª e 119ª. Observe que estas frequências estão na primeira classe, que tem frequência de 1 a 143, então como a tabela é agrupada em classes os valores das posições 1ª a 143ª são considerados iguais ao ponto médio 20.337.

Valor na posição 118ª = 20.337 valor na posição 119ª =20.337

Qi = 20.337× 0,25 + 20.337× 0,75 = 20.337 dólares

Quartil superior: como a posição vale 356,25ª é preciso identificar os valores que estão nas posições 356ª e 357ª. Observe que estas frequências estão na terceira classe, que tem frequência de 329 a 385 (até a segunda classe a frequência acumulada vale 328, então a terceira classe começa na posição 329ª), então como a tabela é agrupada em classes os valores das posições 329ª a 385ª são considerados iguais ao ponto médio 38.685.

Valor na posição 356ª = 38.685 valor na posição 357ª = 38.685

Qs = 38.685 × 0,75 + 38.685 × 0,25 = 38.685 dólares

Como em todas as medidas calculadas a partir de uma tabela agrupada em classes as medidas acima são estimativas dos valores reais. Os valores reais dos quartis (calculados com base nos dados originais do Exemplo 2.8) são:

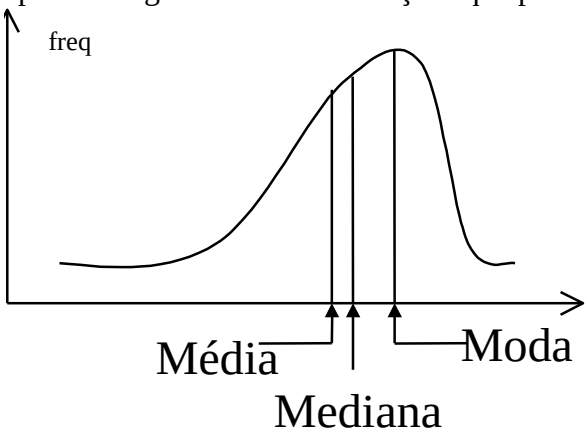
quartil inferior = 24.067,5 dólares *quartil superior = 37.162,5 dólares*
Observe que há uma diferença considerável no quartil inferior (4 mil dólares acima) e um pouco menor no superior (1 mil dólares abaixo).

Tal como nas medidas anteriores é recomendável que sejam usados meios computacionais para o cálculo das separatrizes, sempre de preferência utilizando os dados originais. Conforme mencionado na nota de rodapé 29, os programas computacionais utilizam vários métodos para obtenção dos quartis, quando o número de elementos do conjunto é par. Para os que usarem a planilha eletrônica Microsoft Excel ® para obter os resultados da forma como descritos neste texto deve-se usar a função QUARTIL.EXC:
- quartil inferior - QUARTIL.EXC(intervalo com as células;1)
- mediana (quartil do meio) – QUARTIL.EXC(intervalo com as células;2) ou MED(
- quartil superior – QUARTIL.EXC(intervalo com as células;3)

2.5 - Assimetria das Distribuições

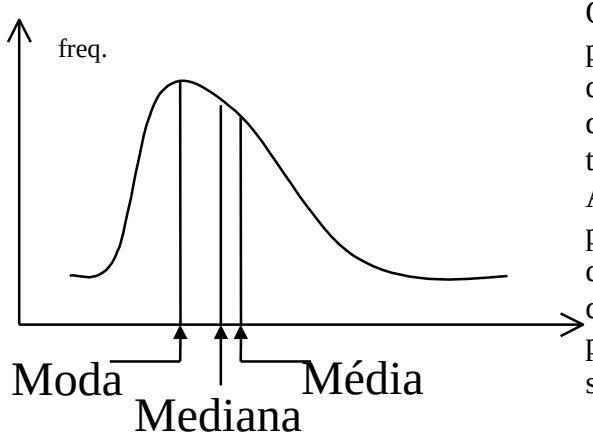
Identificar se a distribuição de uma variável quantitativa em um determinado conjunto de dados é simétrica ou assimétrica pode ser de grande valia por vários motivos:
1) Se os dados são provenientes de uma amostra, identificar a simetria ou não da distribuição pode ser necessário para selecionar o modelo probabilístico mais adequado para descrever a variável na população.
2) No caso de um experimento, em que todas as causas de variação indesejadas são suprimidas, a ocorrência de assimetria quando era esperada simetria, ou o contrário, pode ser indicar que houve algum erro de planejamento ou de medição.
3) Nos casos em que são comparadas distribuições da mesma variável quantitativa em situações diferentes a identificação de um comportamento assimétrico ou simétrico, inesperado ou diferenciado, pode alertar para aspectos anteriormente despercebidos, ou existência de erros.

Alguns programas computacionais calculam uma medida de assimetria (“skewness”): quando este valor é exatamente igual a zero a distribuição em questão é perfeitamente simétrica. Mas a forma ideal de analisar a simetria de uma distribuição é combinar a avaliação das medidas e de um gráfico, seja um histograma ou um diagrama em caixas. Da Figura 25 a Figura 27 irão apresentar gráficos de distribuições que poderiam ser ajustados a histogramas.



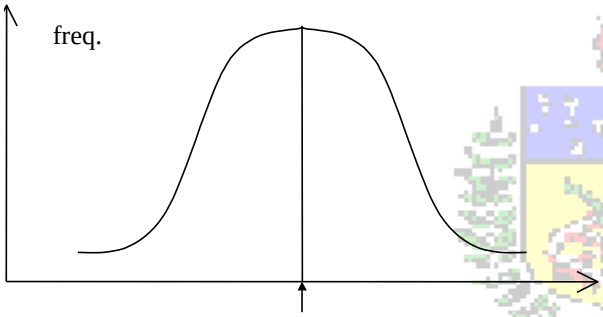
Observe que o "pico" da distribuição, identificado pela moda, está à direita do gráfico, indicando que "falta algo" à esquerda, justificando a denominação "assimétrica à esquerda". Observe também que a mediana é *maior* do que a média. Há uma medida estatística de assimetria que calcula a diferença entre média e mediana: quando a diferença é negativa (mediana maior do que a média) a distribuição é "assimétrica negativa". Este tipo de distribuição não é muito comum na prática, pois é mais difícil obter valores excepcionalmente pequenos (à esquerda)

Figura 25 - Distribuição assimétrica negativa (assimétrica para a esquerda)



Observe que o "pico" da distribuição, identificado pela moda, está à esquerda do gráfico, indicando que "falta algo" à direita, justificando a denominação "assimétrica à direita". Observe também que a média é *maior* do que a mediana. Agora a diferença entre média e mediana será positiva: quando a diferença é positiva a distribuição é "assimétrica negativa". Este tipo de distribuição é razoavelmente comum na prática, pois é fácil obter valores excepcionalmente altos, sendo o caso mais típico a variável renda.

Figura 26 - Distribuição assimétrica positiva (assimétrica para a direita)



Observe que as três medidas de posição coincidem. E que aproximadamente metade dos dados estão abaixo do centro e a outra metade acima, ou seja, a distribuição é "simétrica" em relação às suas medidas de posição. A diferença entre média e mediana é igual a zero. Muitas variáveis apresentam distribuição simétrica, especialmente aquelas resultantes de medidas corpóreas, mas não somente.

Figura 27 - Distribuição simétrica

A seguir apresentamos histogramas de distribuições assimétricas e simétrica.

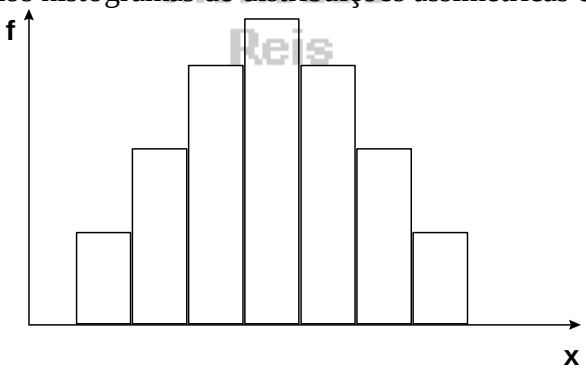


Figura 28 - Histograma de distribuição simétrica

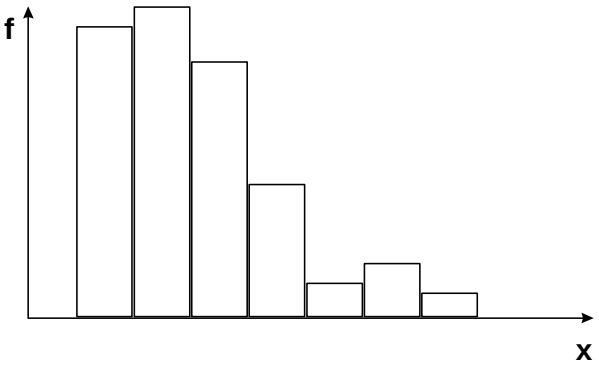


Figura 29 - Histograma de distribuição assimétrica para a direita (negativa)

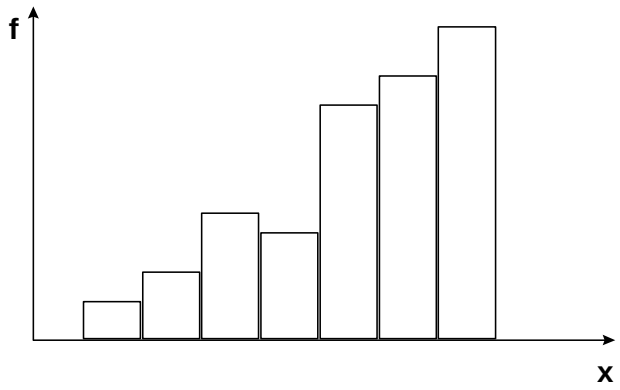
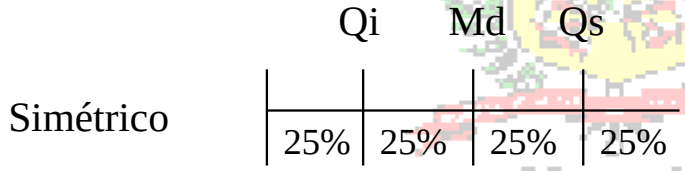


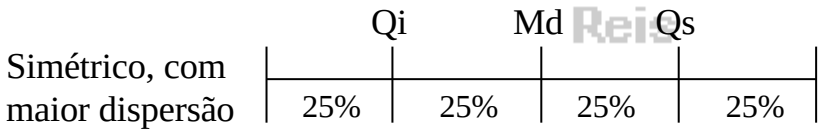
Figura 30 - Histograma de distribuição assimétrica para a esquerda (positiva)

Além das medidas de posição podemos utilizar as separatrizes para avaliar não só a simetria, mas também a dispersão de um conjunto de dados. O procedimento para verificar a existência de assimetria consiste em avaliar a diferença existente entre os quartis e a mediana: se os quartis inferior e superior estiverem à mesma distância da mediana, a distribuição do conjunto pode ser considerada simétrica. A avaliação da dispersão depende da existência de um padrão para comparação, seja um outro conjunto de dados ou alguma especificação. Um conjunto de dados apresentará maior dispersão do que outro se os seus quartis estiverem mais distantes da mediana. Observe as figuras a seguir.



Observe que a diferença Qs - Md é igual à diferença Md - Qi, o que indica a simetria do conjunto. É importante lembrar que os quartis dividem o conjunto em 4 partes iguais (25% dos dados).

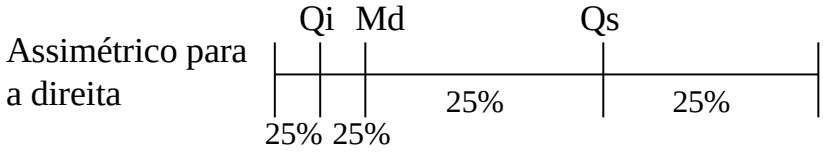
Figura 31 - Quartis de uma distribuição simétrica - 1º caso



Observe que a diferença Qs - Md continua igual à diferença Md - Qi, o que indica a simetria do conjunto.

Figura 32 - Quartis de uma distribuição simétrica - 2º caso

Mas agora a dispersão do conjunto é maior, quando comparada ao 1º caso: os quartis estão mais distantes da mediana (as diferenças Qs - Md e Md - Qi serão maiores do que as obtidas no 1º caso).



Na Figura 33 é fácil perceber diferenças são claramente desiguais: há assimetria. E como Qs - Md é maior do que Md - Qi é para a direita.

Figura 33 - Quartis de uma distribuição assimétrica para a direita

O conjunto apresenta uma dispersão mais elevada nos valores maiores. Isso fez com que o quartil superior aumentasse de valor ("deslocando-o para a direita"), e ficasse mais distante da mediana do que o inferior, significando assimetria para a direita (ou positiva).

Na Figura 34 novamente as diferenças são claramente desiguais: há assimetria. E como Md - Qi é maior do que Qs - Md é para a esquerda.

Figura 34 - Quartis de uma distribuição assimétrica para a esquerda

Neste caso ocorre o oposto da Figura 33. Há maior dispersão nos valores mais baixos, fazendo com que o quartil inferior aumentasse de valor, e ficasse mais distante da mediana do que o superior, significando assimetria para a esquerda (ou negativa).

2.6 - Diagrama em Caixas

O Diagrama em Caixas, também chamado de Desenho Esquemático, Box-plot ou Box & Whisker plot é um gráfico que permite avaliar facilmente os valores típicos, a assimetria, a dispersão e os dados discrepantes de uma distribuição de dados de uma variável QUANTITATIVA. É indicado para grandes conjuntos de dados.

A construção do Diagrama em Caixas exige que sejam calculados previamente os valores da Mediana, Quartil Inferior e Quartil Superior do conjunto de dados, bem como a identificação dos extremos superior (maior valor) e inferior (menor valor). Traçam-se dois retângulos (duas caixas): um representa a "distância" entre o Quartil Inferior e a Mediana e o outro a distância entre a Mediana e o Quartil Superior. A partir dos Quartis Inferior e Superior são desenhadas linhas verticais até os últimos valores não discrepantes tanto abaixo quanto acima.

Valores discrepantes (ou "outliers") são aqueles que têm valores[30]:
- maiores do que a expressão **Qs + 1,5 x (Qs - Qi)** ou
- menores do que a expressão **Qi - 1,5 x (Qs - Qi)**

Todos os valores discrepantes são marcados para posterior estudo individual.

O Diagrama em Caixas "típico" seria:

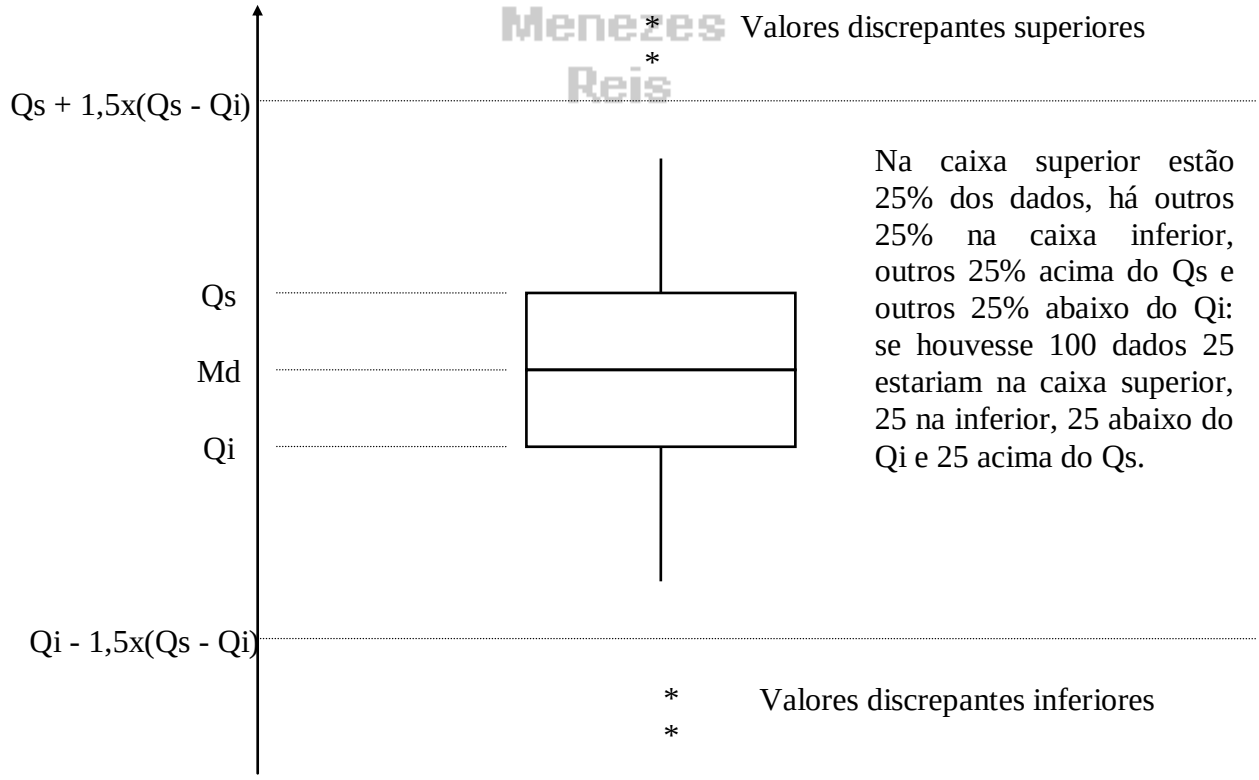


Figura 35 - Diagrama em Caixas - Esquema

[30] O valor **Qs - Qi** é chamado de desvio interquartílico.

Se as duas caixas tiverem "alturas" semelhantes (Qs - Md ≅ Md - Qi) a distribuição é dita simétrica (ver seção 2.4). Quanto maiores as "alturas" das caixas [maiores (Qs - Md) e (Md - Qi)] maior a dispersão do conjunto. O valor "típico" do conjunto será a Mediana (Md), cujas características foram vistas anteriormente. A dimensão horizontal das caixas é irrelevante. Passos para construção do Diagrama em Caixas:

1) Ordenar os dados. 2) Calcular Mediana, Quartil Inferior e Quartil Superior.
3) Identificar Extremos. 4) Construir os retângulos (Qs - Md, Md - Qi).
5) A partir dos retângulos, para cima e para baixo, seguem linhas até o último valor não discrepante.
6) Marcar as observações discrepantes.

Exemplo 2.25 – Construir o diagrama em caixas para os salários anuais dos funcionários da empresa do Exemplo 2.8, avaliando valor típico, assimetria, dispersão e valores discrepantes.

1)Ordenar os dados crescentemente:
Na tabela do Exemplo 2.8 os 474 salários anuais já estavam ordenados.

2) Calcular Mediana, Quartil Inferior e Quartil Superior. Há 474 medidas: ***n = 474***
Posição da mediana = (n + 1) /2 = 475 / 2 = 237,5ª => valor que está na 237,5ª posição
Como o valor foi fracionário, podemos identificar os valores que estão nas posições imediatamente anterior (237ª) e posterior (238ª) e calcular a média entre eles.
Na posição 237ª está o valor 28800 e na 238ª o valor 28950, então:

***Md** = (28800 + 28950)/2 = 28875 dólares*
50% dos funcionários recebem até 28875 dólares por ano e 50% acima disso.

Posição do quartil inferior = (n + 1) / 4 = 475 /4 = 118,75ª => valor que está na 118,75ª posição
Como o valor foi fracionário, podemos identificar os valores que estão nas posições imediatamente anterior (118ª) e posterior (119ª) e usar os critérios vistos na seção 2.4.3:
Na posição 118ª está o valor 24000 e na 119ª o valor 24090, e como a parte decimal da posição vale 0,75:

***Qi** = 24000 × 0,25 + 24090 × 0,75 = 24067,50 dólares*
25% dos funcionários recebem até 24067,50 dólares anuais e 75% acima disso.

Posição do quartil superior = 3 × (n + 1) / 4 = (3×475) /4 = 356,25ª => valor da 356,25ª posição
Como o valor foi fracionário, podemos identificar os valores que estão nas posições imediatamente anterior (356ª) e posterior (357ª) e usar os critérios vistos na seção 2.4.3:
Na posição 356ª está o valor 37050 e na 357ª o valor 37500, e como a parte decimal da posição vale 0,25:

***Qs** = 37050 × 0,75 + 37500 × 0,25 = 37162,50 dólares*
75% dos funcionários recebem até 37162,50 dólares anuais e 25% acima disso.
O desvio interquartílico será: Qs – Qi = 37162,50 – 24067,50 = 13095.
50% dos funcionários recebem entre 24067,50 e 37162,50 dólares anuais.

3) Identificar extremos

O maior valor do conjunto (extremo superior) ***Es** = 135000 dólares*
O menor valor do conjunto (extremo inferior) ***Ei** = 15750 dólares*

4) "Retângulos"

***Qs - Md** = 37162,5 – 28875 = 8287,50 dólares*
***Md - Qi** = 28875 – 24067,50 = 4807,50 dólares*

25% dos funcionários recebem entre 24067,50 e 28875 dólares anuais, e 25% recebem entre 28875 e 37162,50 dólares anuais. A distribuição pode ser considerada ASSIMÉTRICA, pois Qs – Md é DIFERENTE de Md – Qi, e neste caso é uma diferença considerável (Qs – Md é 1,72 vezes maior do que Md – Qi, ou seja, 72% maior).

5) Identificação dos valores discrepantes

Qs – Qi = 37162,50 – 24067,50 = 13095

***1,5 × (Qs - Qi)** = 1,5 × 13095 = 19642,50*

***Qi - 1,5 × (Qs - Qi)** = 24067,50 – 19642,50 = 4425*

Valores menores do que 4425 dólares serão salários anuais discrepantes: como o valor mínimo de salário anual é igual a 15750 dólares, NÃO HÁ valor discrepante inferior no conjunto de dados referente aos salários anuais dos funcionários da Escolástica Ltda.. Assim a linha vertical inferior irá até o último valor não discrepante, que é o mínimo 15750.

***Qs + 1,5 × (Qs - Qi)** = 37162,50 + 19642,50 = 56805*

Valores maiores do que 56805 dólares serão discrepantes: há nada menos do que 52 valores acima de 56805 (ver a tabela no enunciado do Exemplo 2.8), então há 52 valores discrepantes superiores no conjunto de dados referente aos salários anuais dos funcionários da Escolástica Ltda. A linha vertical superior irá até o último valor não discrepante, no caso 56750 dólares (que está na posição 422ª).

Todos os passos anteriores são feitos internamente pelo computador quando se usa um programa estatístico para construir um Diagrama em Caixas, resultando no gráfico a seguir[31]:

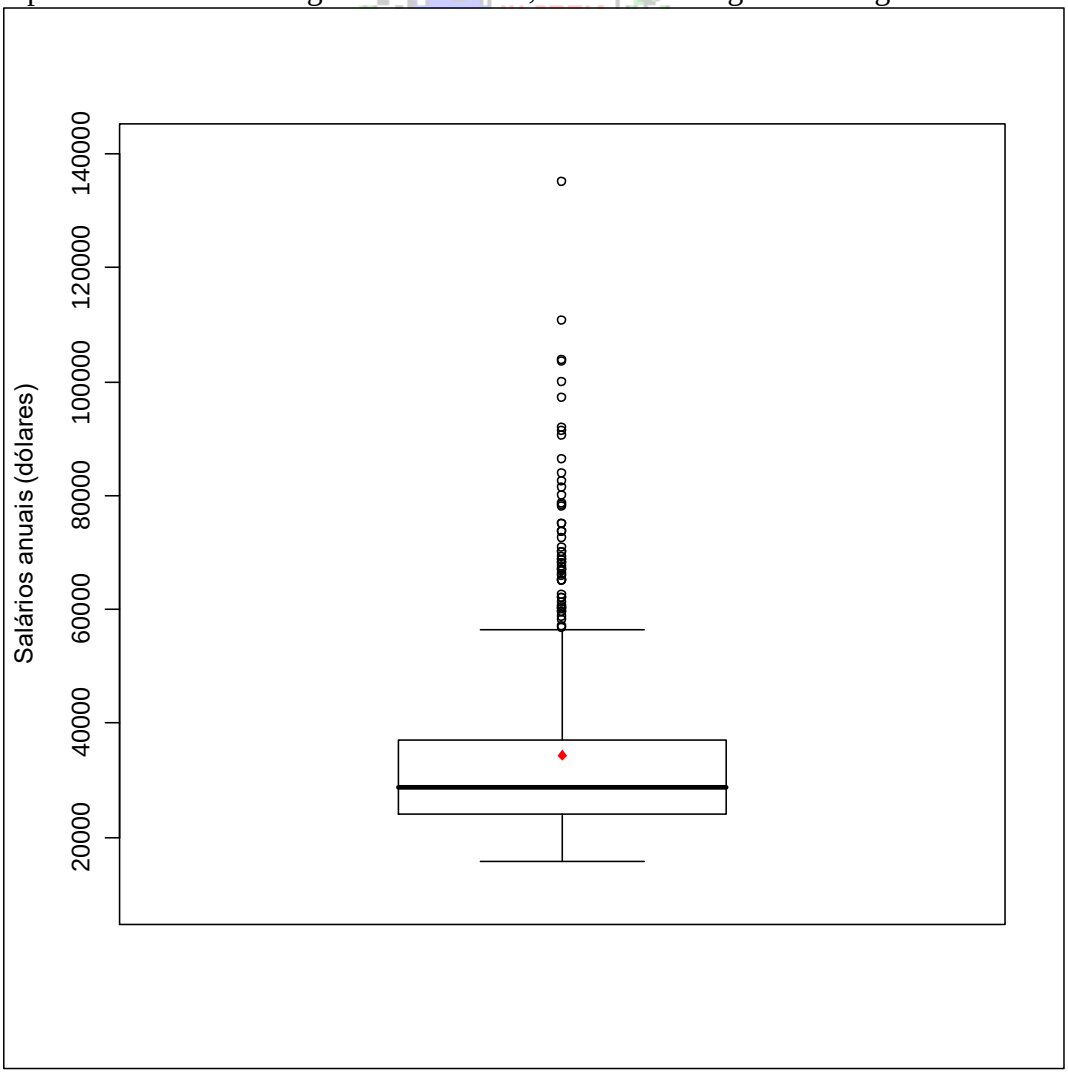


Figura 36 - Diagrama em caixas

Fonte: hipotética

[31] O Diagrama em Caixas foi feito utilizando o pacote Action. Algumas medidas podem ter resultados ligeiramente diferentes dos cálculos manuais devido aos arredondamentos. O ponto marcado é a média dos salários anuais.

O **valor típico** do conjunto é a mediana que vale 28875 dólares. Esse valor pode ser alto ou não, dependendo do objetivo, exigindo conhecimentos mais aprofundados para ser interpretado.

As duas caixas têm "alturas" diferentes (8287,50 e 4807,50 dólares), indicando **assimetria**. Como à distância entre Qs e Md é maior do que a entre Md e Qi trata-se de uma assimetria à direita (ver Figura 33). Quanto à **dispersão** não há o que comentar, pois não há um padrão para comparação.

Não há valores **discrepantes** inferiores, mas há uma grande quantidade de superiores culminando com o valor máximo 135 mil dólares. Estes valores talvez merecessem um estudo individual: primeiramente verificar se não houve erro de registro, se constatada a correção da medida identificar os indivíduos, avaliar as funções exercidas, etc[32].

Como TODA ferramenta estatística o Diagrama em Caixas de nada vale se o usuário não tiver conhecimentos específicos sobre a variável retratada para interpretar os resultados.

2.6.1 - Diagrama em Caixas Múltiplo

É bastante comum querer comparar vários conjuntos de dados, para avaliar seus valores típicos, dispersão, assimetria, e valores discrepantes. Por exemplo, no caso do Exemplo 2.25 poderíamos ter interesse em comparar vários conjuntos de salários, provenientes de diferentes grupos, definidos em função do sexo ou da função dos funcionários da Escolástica Ltda. Para tanto precisamos construir um diagrama múltiplo, em que todos tenham a mesma escala, para possibilitar a comparação (diversos programas estatísticos permitem fazer isso diretamente, ou classificando os dados em função de uma variável independente – de agrupamento).

Exemplo 2.26 - O diagrama em caixas múltiplo na Figura 37 apresenta os salários anuais dos funcionários da empresa Escolástica Ltda. em função do sexo. Faça a análise dos diagramas: valor típico, dispersão, assimetria, valores discrepantes. Existem diferenças significativas entre os sexos quanto ao salário?
Quanto aos valores típicos (medianas) percebe-se a diferença entre os sexos: para o feminino a mediana está em torno de 24000 dólares e no masculino é mais alta, em torno de 32000 dólares.

Quanto à assimetria, os salários dos dois grupos apresentam assimetria, pois as caixas têm "alturas" diferentes. Mas é mais pronunciada entre os homens.

Quanto à dispersão, é maior nos salários masculinos, pois suas caixas são maiores (Quartis mais distantes da Mediana).

Há discrepantes nos dois conjuntos, mas apenas superiores, 11 no feminino e 10 no masculino.

Observa-se que o quartil inferior dos salários masculinos é maior do que o quartil superior dos femininos: enquanto apenas 25% das mulheres recebem salários acima de cerca de 28000 dólares, 75% dos homens têm salários acima deste valor. Ou seja, considerando apenas a categorização por sexo existem diferenças entre os salários recebidos por homens e mulheres, com vantagem para os homens. Contudo, seria recomendável analisar os salários em função de outras variáveis existentes na base de dados: TALVEZ as diferenças nos salários devam-se às funções, anos de estudo, experiência prévia, e não apenas ao sexo dos funcionários.

[32] A detecção de fraudes ou "patrimônios incompatíveis com a renda".

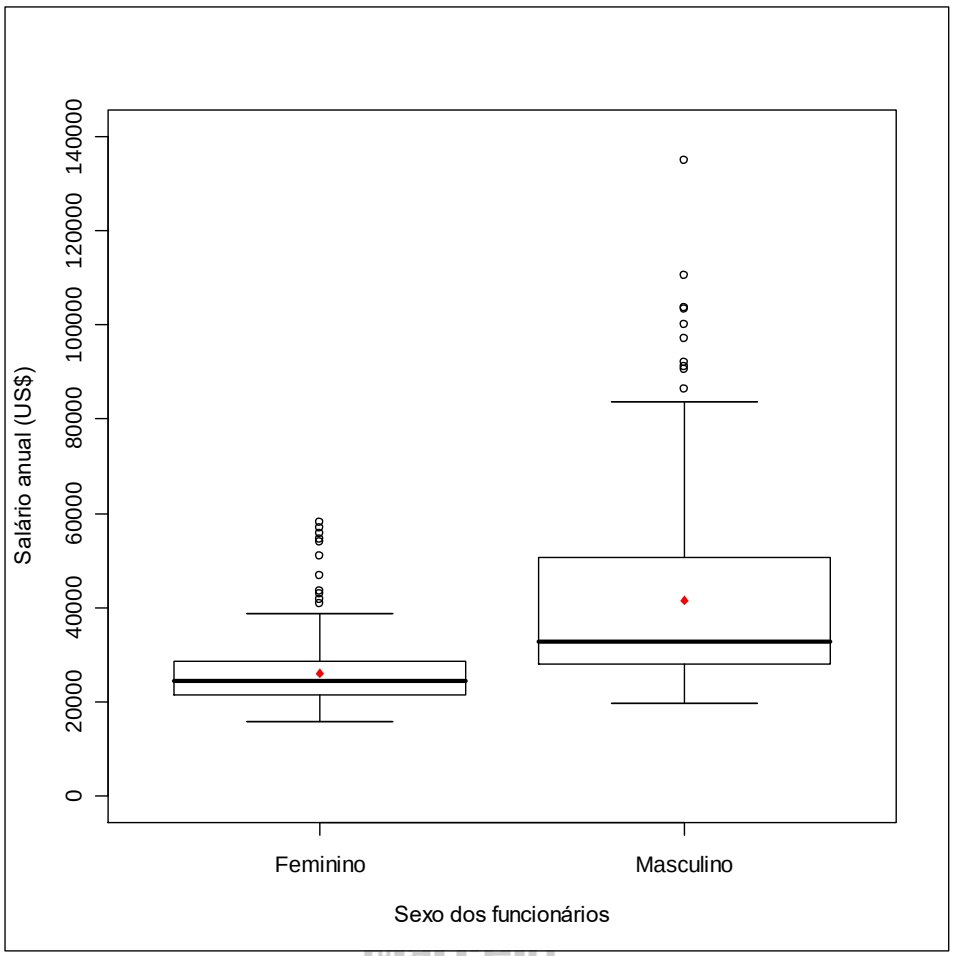


Figura 37 - Diagrama em caixas múltiplo de salários por sexo dos funcionários da empresa Escolástica Ltda.

Exemplo 2.26 - A ONU registrou os crescimentos demográficos e médias de calorias diárias ingeridas em vários países. Os países foram agrupados em seis regiões: OECD (EUA, Canadá, Austrália, Nova Zelândia e Europa Ocidental), África, América Latina, Oriente Médio, Europa Oriental, e Pacífico/Ásia. Os diagramas em caixa das variáveis estão abaixo. Faça a análise dos dois diagramas no que tange aos valores típicos, assimetria, dispersão e valores discrepantes. Qual é a sua opinião sobre a qualidade de vida nestas seis regiões?

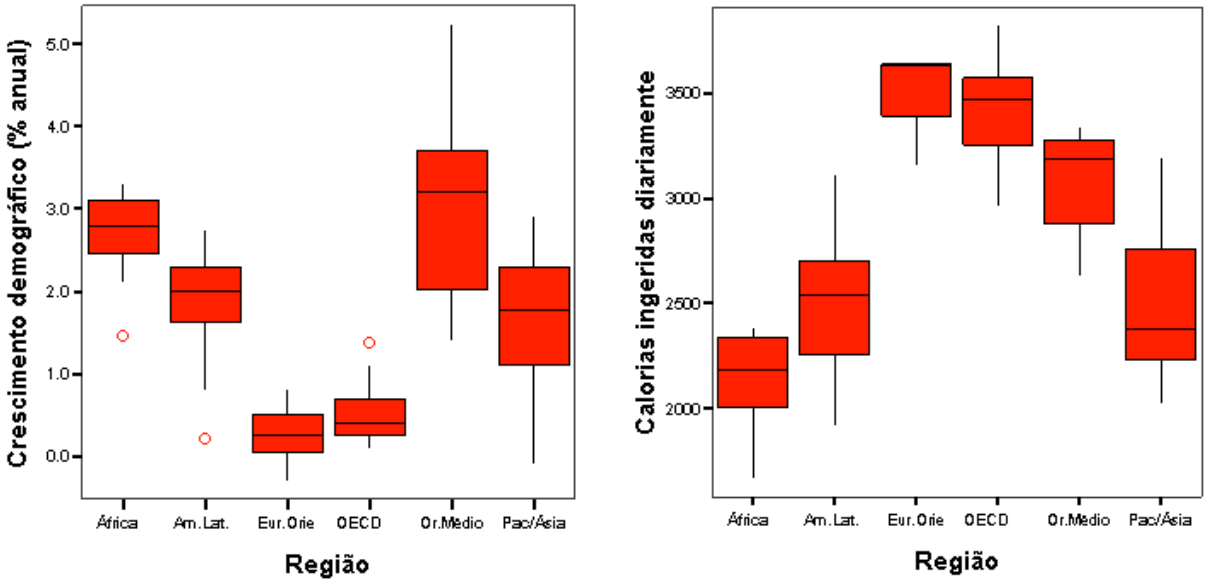


Figura 38 - Diagramas em caixa múltiplos: crescimento demográfico e média diária de calorias ingeridas

Fonte: ONU, 1995

Crescimento demográfico
Valores típicos: Oriente Médio e África têm os maiores valores típicos, medianas de cerca de 3,0% ao ano. E os menores estão na Europa Oriental e OECD, próximos de zero.
Assimetria: os conjuntos de África e Europa Oriental poderiam ser considerados simétricos, América Latina, OECD e Pacífico/Ásia ligeiramente assimétricos, e o Oriente Médio é assimétrico.
Dispersão: o conjunto com maior dispersão é o Oriente Médio, e os menos dispersos são a Europa Oriental e OECD (demonstrando certa homogeneidade demográfica nestas duas regiões).
Valores discrepantes: África e América Latina têm discrepantes inferiores, OECD tem um superior, e as demais regiões não apresentam valores discrepantes.

Média de calorias
Valores típicos: Europa Oriental e OECD têm os maiores valores, na faixa de 3500 calorias diárias, enquanto que a África tem o menor valor, por volta de 2200.
Assimetria: todos os conjuntos são assimétricos, mas Oriente Médio, Pacífico/Ásia e Europa Oriental (onde Qs = Md) são mais do que os outros, a África tem a menor assimetria.
Dispersão: Europa Oriental apresenta a menor dispersão ("caixas" menores), enquanto Pacífico/Ásia apresenta a maior.

É interessante observar o contraste entre os dois diagramas: a África tem um dos maiores valores típicos de crescimento demográfico, e o menor valor típico de calorias ingeridas (indicando um cenário de miséria e fome), enquanto a Europa Oriental e a OECD têm uma situação inversa (o que indica condições sócio-econômicas mais favoráveis). Impressiona também a alta taxa de crescimento demográfico no Oriente Médio.

2.7 – Escolhendo a melhor técnica para analisar os dados

Até o momento estudamos as técnicas de análise exploratória de dados mais importantes. Mas como escolher a técnica que permite melhor organizar e resumir os dados, de maneira que o objetivo do estudo seja alcançado? Diversos fatores vão influenciar na escolha da técnica:
- número de variáveis envolvidas; nível de mensuração das variáveis; objetivo do estudo; tamanho do conjunto de dados; tempo disponível para apresentação dos resultados; público alvo dos resultados do estudo.

2.7.1 – Número de variáveis envolvidas

Se houver mais de uma variável envolvida no estudo é preciso identificar qual (ou quais) é (são) a(s) independente(s) e qual (ou quais) é (são) a(s) dependente(s). É bastante comum haver várias independentes e apenas uma dependente: sendo assim, a descrição dos valores da dependente deve ser feita ***em função*** dos valores das independentes, resultando então em tabelas ou gráficos múltiplos, ou em um "breakdown" das medidas de síntese da variável dependente (medidas para cada grupo formado pelos valores das variáveis independentes).

No Exemplo 2.25 temos duas variáveis envolvidas: sexo e salário anual dos funcionários. A hipótese de estudo é que os salários anuais ***podem*** ser influenciados pelo sexo (funcionários do sexo masculino ***poderiam*** ter salários maiores do que os do sexo feminino, ou o contrário). Então, conclui-se que sexo é a variável independente e salário anual a dependente. Para verificar se a hipótese de estudo é verdadeira optou-se por apresentar os valores da variável dependente em função dos da variável independente, através de um diagrama em caixas múltiplo.

2.7.2 – Nível de mensuração das variáveis

Usualmente queremos descrever o comportamento da variável dependente, portanto torna-se imprescindível utilizar uma técnica apropriada para o seu nível de mensuração. É importante ressaltar que o nome da variável nem sempre é indicação suficiente para identificar o nível de mensuração: velocidade medida em km/h é quantitativa contínua, mas medida como "alta", "média" e "baixa" é qualitativa ordinal. É preciso saber como a variável está sendo avaliada: se através de uma escala de atributos, ela é qualitativa, se através de números (contados ou mensurados por um instrumento de medida) é quantitativa.

É totalmente inadequado utilizar, por exemplo, medidas de síntese, para descrever o comportamento da variável procedência de uma pessoa (sendo esta registrada como o estado natal): como calcular a média entre Acre, São Paulo e Santa Catarina? Da mesma forma, representar os valores da variável velocidade em km/h (que pode assumir teoricamente uma infinidade de resultados) por meio de um gráfico de setores não será apropriado: o gráfico terá um número tão grande de "fatias" que virtualmente não haverá possibilidade de interpretação coerente. A solução possível seria apresentar um gráfico de barras ou setores para a procedência, e um histograma agrupado em classes (ou calcular medidas de síntese, ou um diagrama em caixas) para a velocidade. Aliás, a utilização de tabelas/histogramas agrupados em classes é especialmente útil para variáveis quantitativas ***contínuas***, que podem assumir muitos valores, o que tornaria as tabelas não agrupadas muito extensas e quase que inúteis para analisar os resultados.

2.7.3 – Objetivo do estudo

Obviamente este é um fator crucial. Dependendo do grau de detalhamento que se deseja podemos usar técnicas mais ou menos sofisticadas. Isto é especialmente importante para variáveis quantitativas: o cálculo de média e desvio padrão proporciona um sumário do conjunto de dados, mas um diagrama em caixas possibilita avaliar outros aspectos com maior detalhe. Se há interesse em mostrar o comportamento de uma variável quantitativa ao longo do tempo o gráfico de linhas pode ser uma boa opção. Além disso, conforme visto no item 2.7.1, a hipótese de pesquisa (por exemplo, constatar o efeito de uma variável independente em outra dependente) pode nos obrigar a utilizar algum tipo de técnica: diagrama em caixas múltiplo, um gráfico de setor para cada valor da variável independente, tabulação cruzada, entre outras.

2.7.4 – Tamanho do conjunto de dados

Especialmente importante para variáveis quantitativas. Conforme visto na seção 2.2.2, algumas técnicas que resumem conjuntos de dados referentes a uma variável quantitativa são mais apropriadas para pequenos conjuntos de dados (até 100 observações): diagrama de pontos, ramo-e-folhas, rol. Outras são mais apropriadas para grandes conjuntos, como o histograma e o diagrama em caixas. As medidas de síntese podem ser usadas para qualquer tamanho de conjunto de dados, mas apresentam maior robustez (menor sensibilidade a valores discrepantes e, portanto, representam melhor o comportamento da variável) para grandes conjuntos de dados.

2.7.5 – Tempo disponível para apresentação dos resultados

Atribui-se a Napoleão Bonaparte a seguinte frase: "uma imagem vale mais do que mil palavras". Se não dispomos de muito tempo para apresentar os resultados, em um congresso científico ou em uma reunião de negócios, por exemplo, a utilização de um gráfico apropriado

poderá nos poupar muito tempo: ao invés de ler todas as linhas e colunas de uma complexa tabela o público poderá apreender rapidamente o comportamento da variável e prestar atenção na exposição.

2.7.6 – Público alvo dos resultados do estudo

O conhecimento que o público alvo tem sobre Estatística orientará às vezes decisivamente a escolha da técnica. Devido à veiculação constante pela mídia a maioria das pessoas está familiarizada com gráficos em colunas, setores e linhas, além de tabelas em geral. Dentre as medidas de síntese a única conhecida do grande público é a média, embora constantemente interpretada de forma incorreta. A interpretação do histograma (que nada mais é do que um gráfico de barras justapostas) e do diagrama de pontos também não exigem grande conhecimento estatístico. Contudo, medidas de síntese como mediana, desvio padrão, quartis, coeficiente de variação percentual, e gráficos como ramo-e-folhas e diagrama em caixas somente podem ser interpretados por pessoas que tenham pelo menos tido um curso básico de Estatística, o que não é regra geral.

A Figura 39 sintetiza esta seção, dando especial atenção às técnicas utilizadas para variáveis quantitativas.

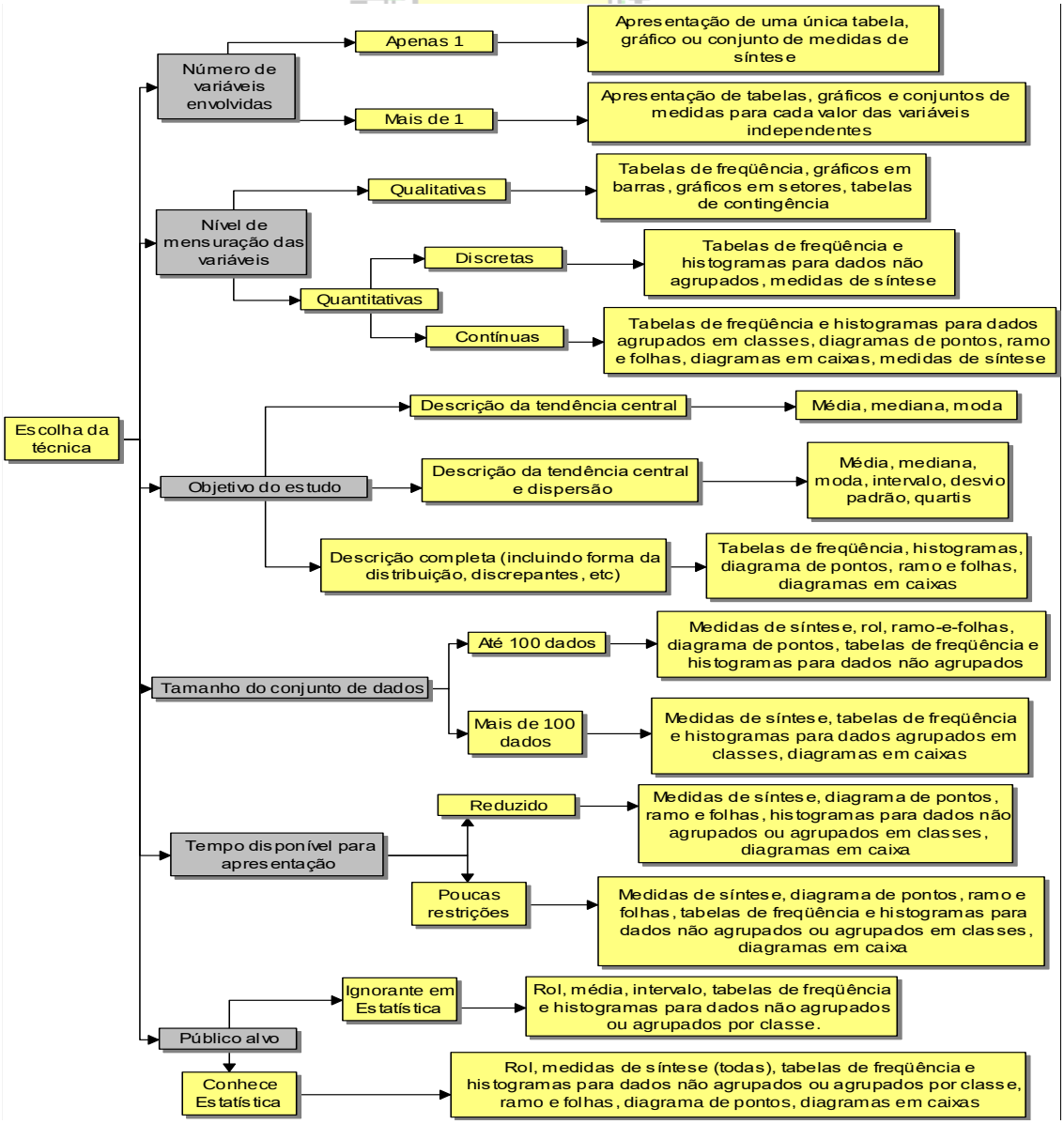


Figura 39 – Fatores que influenciam na escolha de uma técnica de Análise Exploratória de Dados