UNIDADE

3

Análise Exploratória de Dados I

Objetivo

Nesta Unidade, você vai compreender a definição de Análise Exploratória de Dados e aprenderá como realizar a descrição tabular e gráfica de conjuntos de dados referentes a variáveis qualitativas e quantitativas.

O que é Análise Exploratória de Dados?

Caro estudante, nas Unidades anteriores estudamos o planejamento de uma pesquisa e as principais técnicas de amostragem. Conforme vimos, independente de os dados terem sido coletados via censo ou amostragem, eles precisam ser interpretados para atingir os objetivos propostos da pesquisa. O passo inicial para isso é usar os conceitos e técnicas da Análise Exploratória de Dados para resumir e organizar os dados, de maneira que seja possível identificar padrões e elaborar as primeiras conclusões a respeito da população, isto é, descrever a sua variabilidade.

O primeiro passo da Análise Exploratória de Dados é organizar os dados, para que seja possível resumi-los e, posteriormente, interpretá-los. Para entender esse contexto, é importante relembrar a definição de variável e a sua classificação por nível de mensuração e nível de manipulação, estudadas na Unidade 1.

Ainda nesta Unidade, vamos estudar como realizar a análise exploratória de dados através de tabelas e gráficos para cinco casos, tipos de conjuntos de dados: uma variável qualitativa, uma variável quantitativa, duas variáveis qualitativas, uma qualitativa e uma quantitativa, e duas quantitativas.

É indispensável que o administrador seja capaz de realizar Análise Exploratória de Dados: sem isso, a sua capacidade de tomada de decisões ficará seriamente comprometida.

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).

No passado, a Análise Exploratória de Dados era chamada de Estatística Descritiva, por preocupar-se com a descrição dos dados tão-somente.

GLOSSÁRIO

*Variáveis estatísticas – são características que podem ser observadas ou medidas em cada elemento pesquisado, sob as mesmas condições. Para cada variável, para cada elemento pesquisado, em um dado momento, há um e apenas um resultado possível. Fonte: Barbetta (2006).

A partir dessa interpretação inicial, é possível identificar se os dados seguem alguns modelos conhecidos, que permitam estudar o fenômeno sob análise, ou se é necessário sugerir um novo modelo. Usualmente, a concretização dos objetivos de uma pesquisa passa pela análise de uma ou mais **variáveis estatísticas***, ou do seu relacionamento.

O processo da Análise Exploratória de Dados consiste em organizar, resumir e interpretar as medidas das variáveis da melhor maneira possível. Para tanto, é necessário construir um arquivo de dados, que tem algumas características especiais.

Estrutura de um arquivo de dados

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, sejam oriundos de pesquisa de campo ou apenas registros de operações financeiras, arquivos de recursos humanos, entre outros. Possuem 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 (os elementos da população), nas linhas. As variáveis são as características pesquisadas ou registradas. Imagine a base de dados do Departamento de Administração Escolar (DAE) da Universidade Federal de Santa Catarina (UFSC), que armazena as informações dos acadêmicos, contendo as variáveis nome do aluno, data de nascimento, número de matrícula, Índice de Aproveitamento Acumulado (IAA), Índice de Aproveitamento Corrigido (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 2003, supermercado.

Exemplo 1: a Megamontadora Toyord regularmente conduz pesquisas de mercado com os clientes que compraram carros zero km diretamente de suas concessionárias. O objetivo é avaliar a satisfação dos clientes em relação aos diferentes modelos, seu design, adequação ao perfil do cliente. A última pesquisa foi terminada em julho de 2007: 250 clientes foram entrevistados entre o total de 30.000 que compraram veículos novos entre maio de 2006 e maio de 2007. A pesquisa foi restringida aos modelos mais vendidos e que já estão no mercado há dez anos. As seguintes variáveis foram obtidas:

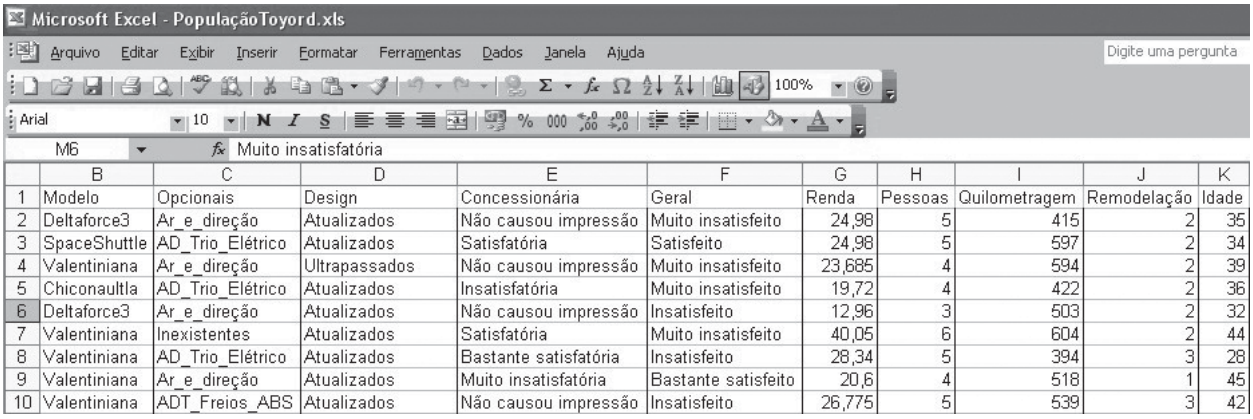
Trata-se de uma empresa fictícia e de uma pesquisa fictícia.

- **modelo comprado**: o compacto Chiconaultla, o sedã médio DeltaForce3, a perua familiar Valentiniana, a van SpaceShuttle ou o luxuoso LuxuriousCar;
- **opcionais**: inexistentes (apenas os itens de série); ar-condicionado e direção hidráulica; ar-condicionado, direção hidráulica e trio elétrico; ar-condicionado, direção hidráulica, trio elétrico e freios ABS;
- **opinião sobre o design**: se os clientes consideram o design do veículo comprado ultrapassado, atualizado ou adiante dos concorrentes;
- **opinião sobre a concessionária onde comprou o veículo (incluindo atendimento na venda, manutenção programada e eventuais problemas imprevistos)**: muito insatisfatória, insatisfatória, não causou impressão, satisfatória, bastante satisfatória;
- **opinião geral sobre o veículo adquirido**: muito insatisfeito, insatisfeito, satisfeito, bastante satisfeito;
- **renda declarada pelo cliente**: em salários mínimos mensais;
- **número de pessoas** geralmente transportadas no veículo;
- **quilometragem** mensal média percorrida com o veículo;

- **percepção do cliente** de há quantos anos o veículo comprado teve a sua última remodelação de design: em anos completos (se há menos de um ano o entrevistador anotou zero); e
- **idade do cliente** em anos completos.

Imagine que você é *trainee* da Toyord. Sua missão é analisar os resultados da pesquisa apresentando um relatório. Dependendo do seu desempenho, você poderá ser contratado em definitivo ou dispensado (sem carta de recomendação). Como deve ser estruturada a base de dados para permitir a análise?

Digamos que você dispõe dos 250 questionários que foram aplicados e você vai tabulá-los em uma planilha eletrônica, como o Microsoft Excel®. Há dez variáveis, a base de dados deve ter, então, dez colunas e 250 linhas (no Excel , 251, já que a primeira será usada para pôr o nome das variáveis). Veja o resultado, com as primeiras linhas (casos), na Figura 16:



	B	C	D	E	F	G	H	I	J	K
1	Modelo	Opcionais	Design	Concessionária	Geral	Renda	Pessoas	Quilometragem	Remodelação	Idade
2	Deltaforce3	Ar_e_direção	Atualizados	Não causou impressão	Muito insatisfeito	24,98	5	415	2	35
3	SpaceShuttle	AD_Trio_Elétrico	Atualizados	Satisfatória	Satisfeito	24,98	5	597	2	34
4	Valentiniana	Ar_e_direção	Ultrapassados	Não causou impressão	Muito insatisfeito	23,685	4	594	2	39
5	Chiconaultla	AD_Trio_Elétrico	Atualizados	Insatisfatória	Muito insatisfeito	19,72	4	422	2	36
6	Deltaforce3	Ar_e_direção	Atualizados	Não causou impressão	Insatisfeito	12,96	3	503	2	32
7	Valentiniana	Inexistentes	Atualizados	Satisfatória	Muito insatisfeito	40,05	6	604	2	44
8	Valentiniana	AD_Trio_Elétrico	Atualizados	Bastante satisfatória	Insatisfeito	28,34	5	394	3	28
9	Valentiniana	Ar_e_direção	Atualizados	Muito insatisfatória	Bastante satisfeito	20,6	4	518	1	45
10	Valentiniana	ADT_Freios_ABS	Atualizados	Não causou impressão	Insatisfeito	26,775	5	539	3	42

Figura 16: Base de dados da Toyord
Fonte: adaptada pelo autor de Microsoft

Veja que cada uma das variáveis é registrada em uma coluna específica e que nas linhas se encontram os registros de cada funcionário. Por exemplo, o respondente 1 adquiriu um modelo DeltaForce3, com os opcionais ar-condicionado e direção hidráulica, considera o

design do veículo atualizado, diz que o atendimento da concessionária onde comprou o veículo não causou impressão, está muito insatisfeito com seu veículo, tem renda mensal de 24,98 salários mínimos (R$ 9.492,00), costuma levar cinco pessoas no veículo, trafega em média 415 km por mês com este veículo, crê que a última remodelação foi feita há dois anos e tem 35 anos de idade. Esse raciocínio pode ser estendido para os outros 249 respondentes. Analisando as variáveis isoladamente ou em conjunto, podemos atingir os objetivos da pesquisa.

O arquivo de dados mostrado na Figura 16 está disponível no Ambiente Virtual de Ensino-Aprendizagem. Juntamente com ele, está disponibilizado o texto "Como realizar análise exploratória de dados no Microsoft Excel".

Veja a seção Saiba mais desta Unidade. O arquivo de dados e o texto servirão para as Unidades 3 e 4.

A grande 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 16. Podemos ter tantas colunas e linhas quantas se quiser, respeitando, porém, as capacidades dos programas. O Microsoft Excel®, por exemplo, admite 65.000 linhas, o que é suficiente para muitas aplicações.

Uma vez os dados no formato apropriado, especialmente se em meio digital, podemos passar para a etapa de análise. Uma das ferramentas mais úteis para isso é a distribuição de freqüências, como veremos a seguir.

Distribuição de freqüências

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

A construção da distribuição de freqüências exige que os possíveis valores da variável sejam discriminados e seja contado o número

GLOSSÁRIO

*Distribuições de freqüências – organizações dos dados de acordo com as ocorrências dos diferentes resultados observados. Fonte: Barbetta, Reis e Bornia (2004).

GLOSSÁRIO

*Freqüência absoluta – registro dos valores da variável por meio de contagem das ocorrências no conjunto de dados. Fonte: Barbetta, Reis e Bornia (2004).

*Freqüência relativa ou percentual – registro dos valores da variável por meio de proporção (relativa) ou percentagem (percentual) do total das ocorrências do conjunto de dados. Fonte: Barbetta, Reis e Bornia (2004).

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 freqüências pode ser expressa através de tabelas ou de gráficos, que terão algumas particularidades dependendo do nível de mensuração da variável e de quantas variáveis serão analisadas. Vamos ver cinco casos: quando há apenas uma variável qualitativa, quando há apenas uma variável quantitativa, quando há duas variáveis (sendo ambas qualitativas, ambas quantitativas, ou uma qualitativa e a outra quantitativa).

Caso de uma variável qualitativa

Usualmente, uma variável qualitativa assume apenas alguns valores: basta, então, discriminá-los e contar quantas vezes eles ocorrem no conjunto. Esta contagem pode ser registrada em números absolutos, **freqüência absoluta***, ou em números relativos, **freqüência relativa ou percentual***. Ambos os registros devem ser feitos e apresentados: a freqüência absoluta permite avaliar se os resultados são sólidos (é temerário tomar decisões com base em pequenas quantidades de dados); já a freqüência relativa possibilita comparar os resultados da distribuição de freqüências com outros conjuntos de tamanhos diferentes. A distribuição de freqüências pode ser apresentada em forma de tabela ou gráfico.

Se alguém diz que 33,33% (percentual) das mulheres de um curso se casaram com professores, você poderia ter uma má impressão destas moças. Mas se alguém diz que das três mulheres (dados brutos) deste curso, uma delas casou-se com um professor, o efeito já não será tão grande. Fonte: Anedota extraída do livro *Como mentir com Estatística*, de Darrel Huff. Rio de Janeiro: Ediouro, 1992.

Exemplo 2: imagine que você está interessado em descrever a variável opinião sobre a concessionária (vista no exemplo 1), isoladamente, e representar os dados em forma de tabela. Como ficariam os resultados? Saiba que o resultado seria semelhante ao ilustrado no Quadro 2, uma apresentação tabular da variável opinião sobre concessionária.

Valores	Freqüência	Percentual
Muito insatisfatória	29	11,60%
Insatisfatória	58	23,20%
Não causou impressão	75	30,00%
Satisfatória	50	20,00%
Bastante satisfatória	38	15,20%
Total	250	100%

Quadro 2: Opinião dos clientes sobre as concessionárias Toyord
Fonte: elaborado pelo autor

Podemos concluir, neste segundo exemplo, que as concessionárias não são exatamente bem-vistas pelos clientes: apenas 35,20% dos entrevistados as consideram satisfatórias ou bastante satisfatórias. Pense que, neste caso, o administrador terá que descobrir as causas de tal resultado e atuar para resolver os problemas.

Podemos aplicar um raciocínio semelhante para as outras variáveis qualitativas e apresentar uma descrição gráfica da distribuição de freqüências. Quando a variável é qualitativa, podemos usar dois tipos de gráficos: **em barras** ou **em setores**.

No gráfico de barras (Figura 17), em um dos eixos são colocadas as categorias da variável, e no outro, as freqüências ou percentuais de cada categoria. As barras podem ser horizontais ou verticais (preferencialmente estas). Para os dados do segundo exemplo, usando as freqüências:

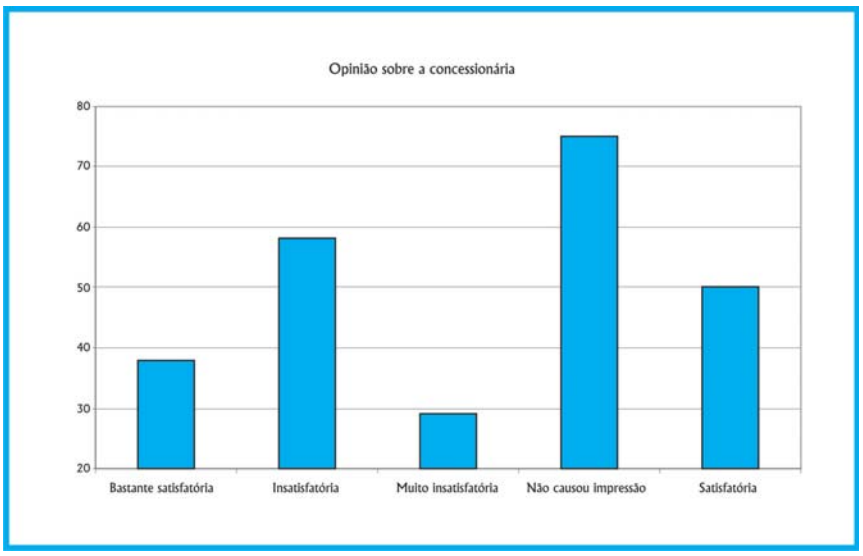


Figura 17: Gráfico em barras para Opinião sobre as concessionárias
Fonte: adaptada pelo autor a partir de Microsoft Office (2007)

Trata-se da mesma distribuição de freqüências observada no Quadro 2. A apreensão da informação, porém, é muito mais rápida através de um gráfico. Percebe-se claramente que a opção "Não causou impressão" apresenta maior freqüência.

Contudo, você consegue identificar alguma particularidade neste gráfico? Olhe bem!

A escala começa em 20, e não em zero. Sendo assim, as diferenças relativas entre as freqüências podem ser distorcidas, o que pode levar a uma interpretação diferente dos resultados: cuidado, portanto, com as escalas dos gráficos. É muito comum vermos erros grosseiros nas escalas de gráficos veiculados na mídia em geral, provavelmente por ignorância, mas devemos estar atentos. Os administradores tomam decisões baseadas na interpretação de gráficos, então estes devem retratar fielmente a realidade. Veja a Figura 18, com a escala correta.

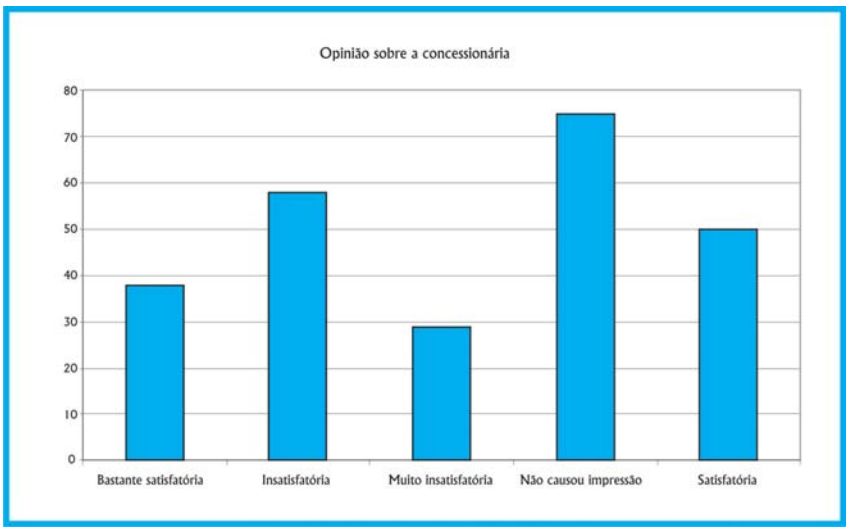


Figura 18: Gráfico em barras para Opinião sobre as concessionárias
Fonte: adaptada pelo autor de Microsoft

Outro tipo de gráfico bastante utilizado é o gráfico circular, em setores ou em "pizza". Ele é apropriado quando o número de valores da variável qualitativa não é muito grande, mas sua construção é um pouco mais elaborada do que o gráfico de barras. Consiste em dividir um círculo (360°) em setores proporcionais às realizações de cada categoria através de uma regra de três simples, na qual a freqüência total (ou o percentual total 100%) corresponderia aos 360°, e a freqüência ou a proporção de cada categoria corresponderia a um valor desconhecido em graus.

$$\textbf{Graus de uma categoria} = \frac{\textbf{360}^{\textbf{o}} \times \textbf{freqüência (proporção) da categoria}}{\textbf{freqüência (proporção) total}}$$

Observe os valores em graus correspondentes aos resultados do Quadro 1 (Quadro 3).

Valores	Freqüência	Percentuais	Graus
Muito insatisfatória	29	11,60%	41,76
Insatisfatória	58	23,20%	83,52
Não causou impressão	75	30,00%	108
Satisfatória	50	20,00%	72
Bastante satisfatória	38	15,20%	54,72
Total	250	100%	360

Quadro 3: Opinião dos clientes sobre as concessionárias Toyord
Fonte: elaborado pelo autor

E o gráfico em setores será conforme apresentado na Figura 19:

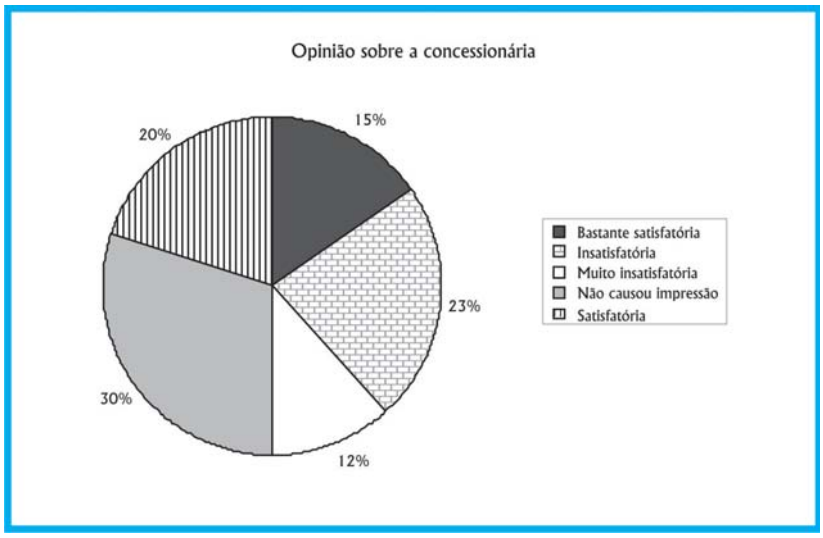


Figura 19: Gráfico em setores para a Opinião sobre as concessionárias
Fonte: adaptada pelo autor de Microsoft®

Pela observação dos percentuais, é possível perceber o predomínio da opção "Não causou impressão" com 30% das respostas. Se a variável qualitativa tiver muitos valores (por exemplo, bairros da região metropolitana de São Paulo), o gráfico dificilmente resumirá alguma coisa, pois terá um número excessivo de fatias. Isso também ocorre com variáveis quantitativas, especialmente as contínuas.

Caso de uma variável quantitativa

A construção das distribuições de freqüências 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 algumas particularidades, dependendo se a variável é **discreta** ou **contínua**.

Se a variável for quantitativa discreta e puder assumir apenas alguns valores, a abordagem será semelhante à das variáveis qualitativas. A diferença reside na substituição de atributos por números, gerando uma **distribuição de freqüência para dados não agrupados**. Vamos ver um exemplo.

Neste terceiro exemplo – para a mesma situação do Exemplo 1 –, imagine que você está interessado em descrever a variável Número de pessoas usualmente transportadas no veículo, isoladamente, e representar os dados em forma de tabela. Como ficariam os resultados? O resultado seria semelhante ao mostrado no Quadro 4, uma apresentação tabular (em forma de tabela) da variável número de pessoas transportadas.

Valores	Freqüência	Percentual
1	19	7,60%
2	29	11,60%
3	43	17,20%
4	42	16,80%
5	57	22,80%
6	60	24,00%
Total	250	100%

Quadro 4: Número de pessoas usualmente transportadas no veículo
Fonte: elaborado pelo autor

Pela observação do Quadro 4, podemos concluir que os veículos têm uso predominantemente "familiar" (várias pessoas transportadas usualmente). Sabendo disso, o administrador pode decidir por direcionar o marketing ou mesmo a produção de modelos visando ao segmento de famílias maiores. Uma abordagem semelhante poderia

ser aplicada para as outras variáveis discretas: anos de remodelação e mesmo idade dos consumidores.

E como representar a distribuição de freqüências para variáveis quantitativas discretas graficamente? O Quadro 3 poderia ser representado através de um histograma, um gráfico de barras justapostas, em que as áreas das barras são proporcionais às freqüências de cada valor. Vamos ver (Figura 20):

A maioria dos programas (estatísticos ou não) que constroem histogramas para variáveis quantitativas discretas costuma ignorar isso.

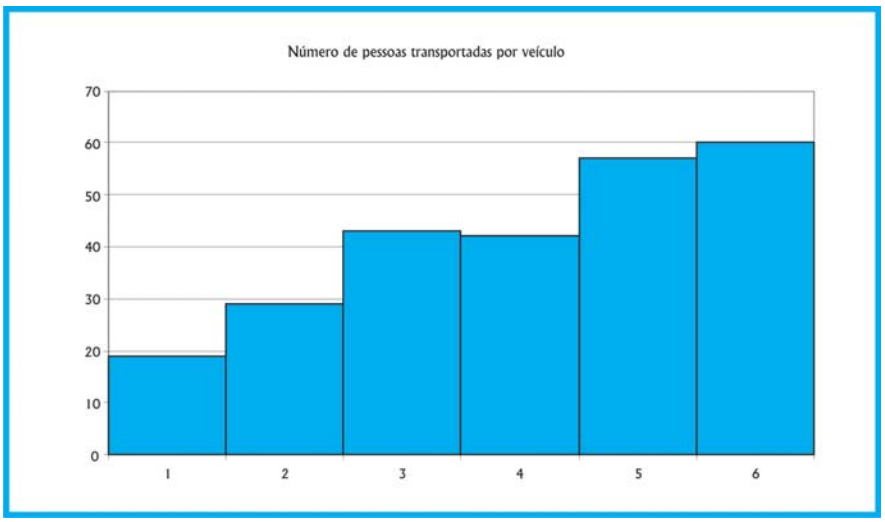


Figura 20: Histograma do Número de pessoas transportadas por veículo
Fonte: adaptada pelo autor de Microsoft

Neste caso, eu poderia usar o gráfico em setores? A resposta é não, pois formalmente o gráfico em setores deve ser usado apenas para variáveis qualitativas. A interpretação é a mesma, mas a apreensão da informação é mais rápida. Observe que não há problemas com a escala vertical, pois esta começa em zero.

Se a variável quantitativa for contínua, o procedimento descrito anteriormente será inviável como instrumento de resumo do conjunto, pois praticamente todos os valores têm freqüência baixa, o que resultaria em uma tabela enorme.

Se o conjunto de dados for pequeno, até cem observações, é possível usar ferramentas gráficas como o diagrama de pontos e o ramo em folhas.

Mais informações, veja a seção Saiba mais.

Se o conjunto for grande, é preciso representar os dados através de um conjunto de faixas de valores mutuamente exclusivas (para que cada valor pertença apenas a uma faixa), que contenha do menor ao maior valor do conjunto: registram-se, então, quantos valores do conjunto se encontram em cada faixa. Há duas maneiras de fazer isso:

- através da **categorização*** (recodificação) da variável, por exemplo, todos que ganham até 4 salários mínimos (R$ 1.520) pertencem à classe baixa, todos que ganham entre 4,01 e 20 salários mínimos (até R$ 7.600) pertencem à classe média, e acima disso pertencem à classe alta – esta abordagem é largamente utilizada na mídia; e
- através de uma **distribuição de freqüências para dados agrupados*** (ou agrupada em classes), processo mais elaborado e mais "estatístico". Veremos o procedimento a seguir.

O processo para montagem da distribuição de freqüências para dados agrupados é o seguinte:

- determinar o intervalo do conjunto (diferença entre o maior e o menor valor do conjunto);
- dividir o intervalo em um número conveniente de classes, onde: $\text{Nº de classes} = \sqrt{\text{Nº de elementos}}$. Neste ponto, há grande controvérsia entre os estatísticos, e a fórmula apresentada é apenas uma das opções possíveis. Admite-se que o número mínimo de classes seja igual a 5, e o máximo, 20, mas se aceita uma definição arbitrária neste intervalo;
- estabelecer as classes com a seguinte notação:

  Li – limite inferior;

  Ls – limite superior;

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

  Li |--| Ls ambos incluídos;
- determinar as freqüências de cada classe; e

GLOSSÁRIO

*Categorização – processo pelo qual se transforma uma variável quantitativa em qualitativa, associando atributos a intervalos de valores numéricos, por exemplo, classe A para uma certa faixa de renda familiar. Fonte: elaborado pelo autor

*Distribuição de freqüências para dados agrupados – distribuição de freqüências na qual os valores da variável são agrupados em faixas de ocorrência, e as freqüências, contadas para cada faixa, para facilitar o resumo do conjunto de dados, usualmente empregado para variáveis quantitativas contínuas. Fonte: Barbetta, Reis e Bornia (2004).

- determinar os pontos médios de cada classe através da média dos dois limites (serão os representantes das classes).

Vamos ver exemplos de ambas as abordagens.

A resolução passo a passo deste problema está na seção Saiba mais desta Unidade, que explica como realizar análise exploratória de dados no Excel®. Aqui apresentaremos apenas os resultados finais.

Exemplo 4: para a mesma situação do Exemplo 1 – imagine que você está interessado em descrever a variável Renda dos consumidores, isoladamente, e representar os dados em forma de tabela. Como ficariam os resultados nos seguintes casos:

a) se optássemos por categorizar a variável da seguinte forma: todos que ganham até 4 salários mínimos (R$ 1.520) pertencem à classe baixa, todos que ganham entre 4,01 e 20 salários mínimos (até R$ 7.600) pertencem à classe média, e acima disso pertencem à classe alta?; e

b) se optássemos por uma distribuição de freqüências para dados agrupados?

No caso do item a, a categorização levará à criação de uma nova variável, agora qualitativa, permitindo uma abordagem semelhante à que vimos anteriormente. No Quadro 5 e na Figura 21, estão os resultados: tabela de freqüências e gráfico em setores.

Valores	Freqüência	Percentual
Classe baixa (até 2 s.m.)	2	0,8%
Classe média (entre 2,01 e 20 s.m.)	104	41,6%
Classe alta (acima de 20 s.m.)	144	57,6%
Total	250	100%

Quadro 5: Renda categorizada em classe social
Fonte: elaborado pelo autor

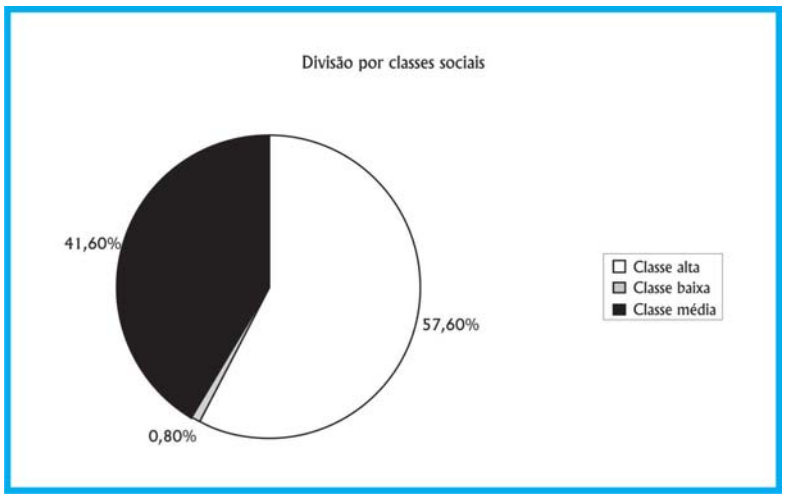


Figura 21: Gráfico em setores para a Renda categorizada em classes
Fonte: adaptada pelo autor de Microsoft Office (2007)

Observe que perdemos informação sobre os dados originais de renda ao fazer a categorização. A interpretação é relativamente simples: a maioria absoluta (mais de 50%) dos clientes da montadora pode ser considerada de classe alta (renda superior a 20 salários mínimos mensais). A grande discussão que surge neste caso é quem define o que é classe baixa, média ou alta (ou A, B, C, D e E). Uma sugestão é utilizar a classificação do IBGE.

Passando para o item b, devemos seguir os passos:

- Intervalo = Maior – Menor = 86,015 – 1,795 = 84,22 (a maior renda é de 86,015 salários mínimos, e a menor, de 1,795, as classes devem englobar do menor ao maior valor);

Estes valores foram obtidos no arquivo de dados citado no início desta Unidade.

- Nº de classes = $\sqrt{\text{Nº de elementos}} = \sqrt{250} = 15{,}81 \cong 16$. Por este expediente, deveríamos usar 16 classes. Porém, conforme foi dito anteriormente, o número de classes pode ser definido de forma arbitrária: para simplificar nosso problema, vamos usar 5 classes.

  Amplitude das classes = 86,015/5 = 16,844 (valor exato)

  A amplitude das classes pode ser ligeiramente maior do que a obtida acima, poderíamos, novamente procurando a simplificação do problema, usar amplitude igual a 16,85. Se a

amplitude não for um valor exato, deve sempre ser arredondada para cima, garantindo que as classes conterão do menor ao maior valor. As classes podem, então, ser definidas;

- Classes: 1,795|-18,645 18,645|-35,495 35,495|-52,345 52,345|-69,195 69,195|-86,045 (neste caso, o ponto inicial foi o próprio menor valor do conjunto, poderia ser outro valor conveniente abaixo do menor valor);
- pontos médios de cada classe: (limite inferior + limite superior)/2 (os pontos médios calculados estão no quadro abaixo); e
- freqüências de cada classe (Quadro 6):

Classes	Freqüência	Percentuais	Pontos médios
1,795|-18,645	98	39,2%	10,22
18,645|-35,495	102	40,8%	27,07
35,495|-52,345	38	15,2%	43,92
52,345|-69,195	9	3,6%	60,77
69,195|-86,045	3	1,2%	77,62
Total	250	100%	-

Quadro 6: Renda agrupada em classes
Fonte: elaborado pelo autor

Observe que perdemos informação sobre o conjunto original: sabe-se que há 98 pessoas com renda entre 1,795 e 18,645 salários mínimos, mas não quais são os seus valores exatos, ou seja, as freqüências das classes passam a ser as freqüências dos pontos médios. Podemos afirmar que quase 80% dos clientes têm renda até 35,495 salários mínimos.

O Quadro 6 também pode ser representado através de um histograma (Figura 22), uma vez que a variável permanece sendo formalmente quantitativa. Mas o histograma para uma tabela de dados agrupados é um pouco diferente do visto anteriormente. O número de barras é igual ao número de classes. Cada barra é centrada no ponto médio de cada classe, o ponto inicial de cada barra é o limite inferior da classe, e o ponto final é o limite superior.

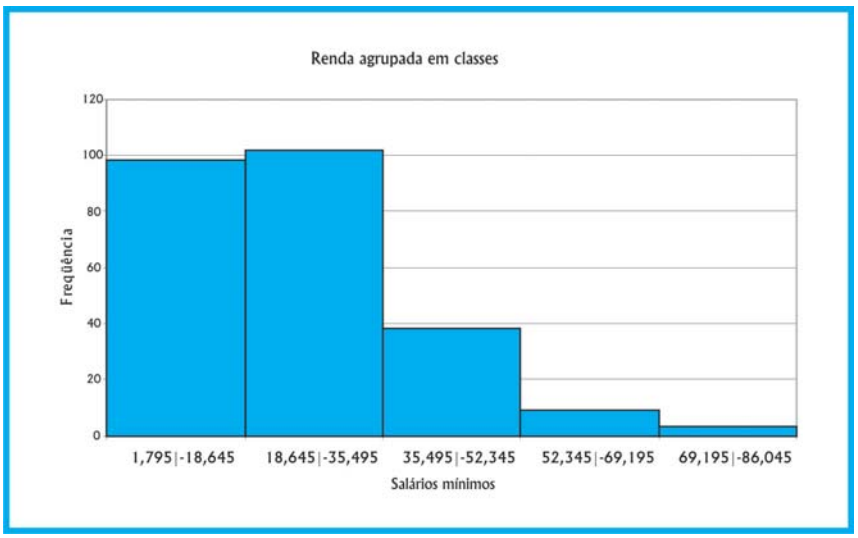


Figura 22: Histograma para Renda agrupada em classes
Fonte: adaptada pelo autor de Microsoft Office (2007)

Note que a interpretação é mais direta quando olhamos o gráfico apresentado na Figura 20.

Mas aqui surge um fato interessante. Parece haver contradição com a interpretação do item a, na qual concluímos que a maioria absoluta dos clientes é de classe alta. Isso ocorre devido à definição arbitrária das classes, e à ainda mais arbitrária definição de classes baixa, média e alta. Pense em como você resolveria esta contradição.

O agrupamento em classes apresenta algumas desvantagens, além da já citada perda de informação sobre o conjunto original.

Os pontos médios nem sempre são os representantes mais fiéis das classes. Para uma grande quantidade de dados, existe uma maior probabilidade de que estas estimativas correspondam exatamente aos verdadeiros valores. Outro problema são as medidas estatísticas calculadas com base na distribuição de freqüências para dados agrupados: serão apenas estimativas dos valores reais devido à perda de informação referida acima.

A tendência atual é NÃO CALCULAR medidas estatísticas com base em tabelas de dados grupados.

Agora, vamos ver como analisar o relacionamento entre duas variáveis. Começaremos com duas variáveis qualitativas.

Caso de duas variáveis qualitativas

O administrador freqüentemente precisa estudar o relacionamento entre duas ou mais variáveis, para tomar decisões. Por exemplo, há relação entre o sexo do consumidor e a preferência por um modelo de carro, ou entre a escolaridade do eleitor e o candidato a presidente escolhido, entre outras.

Quando as duas variáveis são qualitativas (originalmente ou quantitativas categorizadas), usualmente é construída uma distribuição conjunta de freqüências, também chamada de **tabela de contingências*** ou dupla classificação. Nela são contadas as freqüências de cada cruzamento possível entre os valores das variáveis. A expressão pode incluir o cálculo de percentuais em relação ao total das linhas, colunas ou total geral da tabela. A representação gráfica também é possível. Vamos ver um exemplo.

Para a mesma situação do Exemplo 1. Agora, você está interessado em observar o relacionamento entre a variável Modelo adquirido e a Opinião geral do cliente sobre o veículo, e expressá-lo de forma tabular e gráfica.

A variável Modelo apresenta cinco resultados possíveis (cinco modelos foram considerados nesta pesquisa), e a variável Opinião geral pode assumir quatro resultados (bastante satisfeito, satisfeito, insatisfeito e muito insatisfeito). Isso significa que podemos ter até 20 cruzamentos possíveis para os quais precisamos contar as freqüências. Para grandes bases de dados, mesmo para o nosso exemplo em que há apenas 250 casos, seria um processo tedioso e sujeito a erros. Portanto, o mais inteligente é utilizar alguma ferramenta computacional, mesmo uma planilha eletrônica como o Microsoft Excel®.

Usando uma ferramenta computacional, chegaremos ao Quadro 7.

GLOSSÁRIO

*Tabela de contingências – tabela que permite analisar o relacionamento entre duas variáveis; nas linhas, são postos os valores de uma delas, e nas colunas, os da outra, e nas células contam-se as freqüências de todos os cruzamentos possíveis. Fonte: Barbetta (2006).

Opinião geral sobre o veículo					
Modelo	**Muito insatisfeito**	**Insatisfeito**	**Satisfeito**	**Bastante satisfeito**	**Total**
		1			1
Chiconaultla	69	11	1	0	81
DeltaForce3	29	22	5	0	56
Valentiniana	11	18	9	3	41
SpaceShuttle	1	14	17	10	42
LuxuriousCar	0	1	9	19	29
Total	110	67	41	32	250

Quadro 7: Tabela de contingências de modelo por opinião geral (apenas freqüências)
Fonte: elaborado pelo autor

Observe a última coluna e a última linha do quadro acima: são os chamados **totais marginais***, isto é, as freqüências dos valores das variáveis Modelo e Opinião geral sobre o veículo, respectivamente. Percebe-se que os modelos Chiconaultla e DeltaForce3 são os mais vendidos, e que as opiniões negativas (muito insatisfeito e insatisfeito) são mais freqüentes do que as positivas.

Além disso, é fácil perceber que as opiniões negativas são as predominantes nos modelos Chiconaultla, DeltaForce3 e, em menor grau, no Valentiniana. Apenas os modelos SpaceShuttle e LuxuriousCar têm proprietários predominantemente satisfeitos.

Você deve ter percebido também uma linha com várias células vazias (apenas uma observação na opção insatisfeito). Trata-se de um **dado perdido**: o entrevistado esqueceu de mencionar o modelo adquirido, ou o entrevistador não o registrou durante a realização da pesquisa, ou mesmo houve um erro de digitação. Como a quantidade aqui é muito pequena (1 em 250, 0,4%), não causará grandes problemas. Apenas quando a quantidade ultrapassa 5% da base de dados, há motivo para preocupação, pois houve muitos erros de digitação na tabulação dos dados ou o instrumento de pesquisa foi mal projetado, pois muitos elementos da população não forneceram as informações desejadas.

GLOSSÁRIO

*Totais marginais – totais das linhas ou das colunas de uma tabela de contingência, permitem avaliar individualmente as variáveis componentes da tabela. Fonte: Bussab e Morettin (2002).

Conforme vimos na Unidade 1.

O Quadro 7 pode ser apresentado de forma gráfica, através de um gráfico de barras múltiplas (Figura 23).

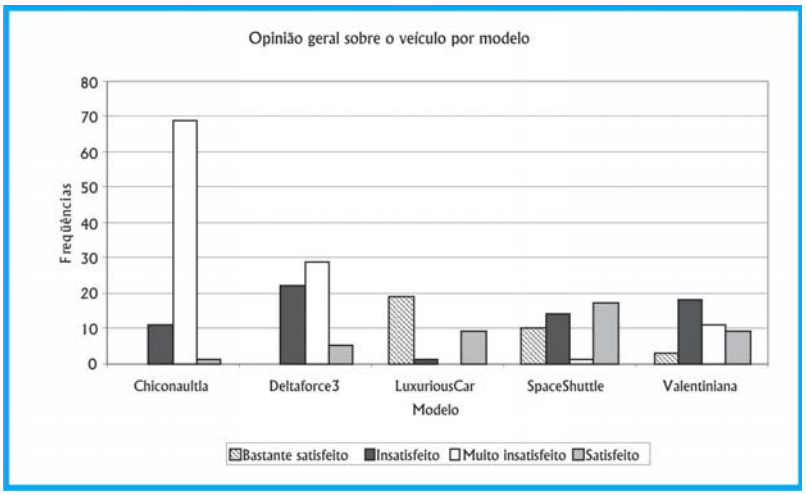


Figura 23: Gráfico de barras múltiplas: Opinião geral por modelo
Fonte: adaptado pelo autor de Microsoft Office (2007)

As freqüências absolutas podem ser insuficientes para a interpretação dos resultados, especialmente quando comparando os resultados com outros conjuntos de dados de tamanhos diferentes. Assim, podemos calcular percentuais em relação aos totais de cada coluna, aos totais de cada linha ou ao total geral da tabela. Vamos apresentar apenas um dos percentuais possíveis, em relação aos totais das linhas. No texto "Como realizar análise exploratória de dados no Microsoft Excel®", são apresentados todos os resultados (Quadro 8):

Opinião geral sobre o veículo									
Modelo	Muito insatisfeito		Insatisfeito		Satisfeito		Bastante satisfeito		Total	
Chiconaultla	69	85,19%	11	13,58%	1	1,23%	0	0,00%	81	100%
DeltaForce3	29	51,79%	22	39,29%	5	8,93%	0	0,00%	56	100%
Valentiniana	11	26,83%	18	43,90%	9	21,95%	3	7,32%	41	100%
SpaceShuttle	1	2,38%	14	33,33%	17	40,48%	10	23,81%	42	100%
LuxuriousCar	0	0,00%	1	3,45%	9	31,03%	19	65,52%	29	100%
Total	110	44,18%	66	26,51%	41	16,47%	32	12,85%	249	100%

Quadro 8: Tabela de contingência de Opinião geral por modelo (com % por linha)
Fonte: elaborado pelo autor

Vistos os exemplos, o que você pode concluir acerca da satisfação dos clientes com relação aos modelos? Qual modelo deveria receber atenção prioritária?

Veja que o cruzamento de duas variáveis qualitativas é atividade corriqueira para o administrador, e cada vez mais esse profissional precisa avaliar mais de duas variáveis, o que exige métodos matemáticos sofisticados, implementados computacionalmente. Veremos mais sobre esse tema a seguir.

Caso de duas variáveis quantitativas

Muitas vezes, também estamos interessados em avaliar o relacionamento entre variáveis quantitativas, sejam elas discretas, sejam contínuas.

Basicamente, há interesse em, a partir de dados, verificar **se** e **como** duas variáveis quantitativas relacionam-se entre si em uma população, ou seja, avaliar se há **correlação*** entre elas, e avaliar a força e a direção (se elas caminham na mesma direção ou em direções opostas) desta correlação, caso ela exista.

Uma das variáveis é chamada de independente. Esta pode ser uma variável que o pesquisador manipulou para observar o efeito em outra ou alguma cuja medição possa ser feita de maneira mais fácil ou precisa, sendo, então, suposta sem erro.

Há uma outra variável, chamada de dependente. Seus valores são resultado da variação dos valores das variáveis independentes.

Reveja as definições de variáveis na Unidade 1.

Esta denominação costuma levar à má interpretação do significado da "correlação" entre variáveis: se há correlação entre variáveis, significa que os seus valores variam em uma mesma direção ou em direções opostas, com uma certa "força", ou seja, correlação não significa causalidade.

GLOSSÁRIO

*Correlação – medida de associação entre duas variáveis quantitativas. Fonte: Barbetta, Reis e Bornia (2004).

Por exemplo, pode haver correlação entre a pluviosidade mensal (em mm) em Florianópolis e o número de ratos exterminados por mês na cidade de Sidney, na Austrália, mas seria um pouco forçado imaginar que uma coisa "causou" a outra. É necessário usar bom senso.

Em outro caso, ao avaliarmos o relacionamento entre renda mensal em reais e área em m^2 da residência de uma família, esperamos um relacionamento positivo entre ambas: para maior renda (independente), esperamos maior área (dependente).

GLOSSÁRIO

*Observações emparelhadas – medidas de duas ou mais variáveis que foram realizadas na mesma unidade experimental/amostral, no mesmo momento. Fonte: elaborado pelo autor

Para que seja possível avaliar o relacionamento entre duas variáveis (neste caso, quaisquer, não apenas quantitativas), os dados devem provir de **observações emparelhadas*** e em condições semelhantes. Ao avaliar a correlação existente entre a altura e o peso de um determinado grupo de crianças, por exemplo, o peso de uma determinada criança deve ser medido e registrado no mesmo instante em que é medida e registrada a sua altura. Renda e área da residência da mesma família, no mesmo momento.

Se estivermos analisando duas variáveis quantitativas, cujas observações constituem pares ordenados, chamando estas variáveis de **X** (independente) e **Y** (dependente), podemos plotar o conjunto de pares ordenados (x,y) em um diagrama cartesiano, que é chamado de **diagrama de dispersão**. Atualmente, isso pode ser feito com aplicativos computacionais, até mesmo uma planilha eletrônica como o Microsoft Excel®.

Saiba mais no texto "Como realizar análise exploratória de dados no Microsoft Excel®".

Através do diagrama de dispersão, podemos ter uma idéia inicial de como as variáveis estão relacionadas: a direção da correlação (isto é, quando os valores de **X** aumentam, os valores de **Y** aumentam também ou diminuem), a força da correlação (em que "taxa" os valores de **Y** aumentam ou diminuem em função de **X**) e a natureza da correlação (se é possível ajustar uma reta, parábola, exponencial, aos pontos).

Vamos a um exemplo para ilustrar.

Para a mesma situação do Exemplo 1, gostaríamos de saber como é o relacionamento entre a variável Renda mensal do cliente e a Quilometragem média mensal por ele percorrida.

As duas variáveis de interesse (Renda e Quilometragem) são quantitativas. Quem pode influenciar quem? É mais lógico imaginar que, quanto maior a renda familiar, haverá mais dinheiro para comprar combustível, e, portanto, maior a quilometragem percorrida com o veículo. Sendo assim, a variável renda será posta no eixo horizontal (X) do diagrama de dispersão, e a quilometragem, no eixo vertical (Y). Veja a Figura 24:

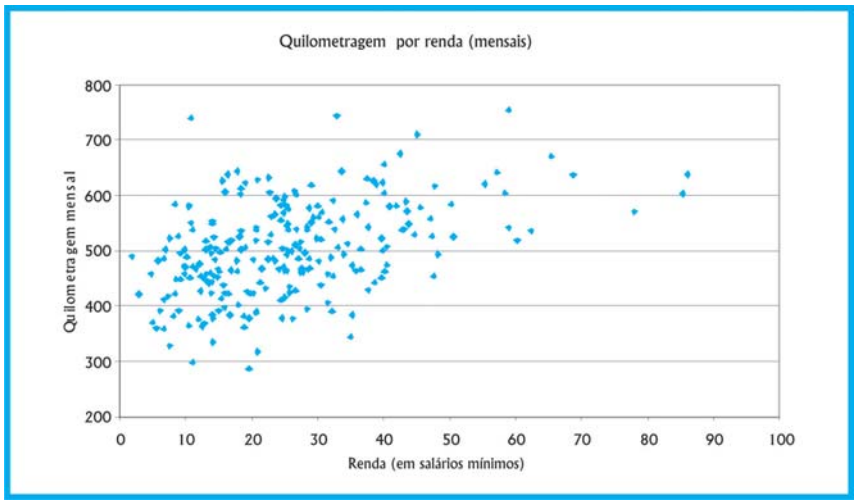


Figura 24: Diagrama de dispersão de Quilometragem por Renda
Fonte: adaptada pelo autor de Microsoft®

Aparentemente, a correlação entre as variáveis é positiva, como era esperado: maiores valores de renda correspondem a maiores valores de quilometragem. A correlação não parece ser muito forte, pois os pontos não estão muito próximos. Quanto à natureza, é difícil afirmar, talvez seja linear, mas é apenas um palpite neste caso.

Para esclarecer, vamos ver outro exemplo.

Neste caso, uma empresa agroindustrial processa soja para obter óleo. A direção quer estudar o relacionamento entre o valor da soja (em dólares por tonelada) na Bolsa de Cereais de Chicago e a cotação da ação da empresa (em dólares) na Bolsa de Nova York. Para tanto,

coletou um conjunto de 400 pares de observações e plotou o diagrama de dispersão exposto na Figura 25.

Observando o diagrama (Figura 25), é possível afirmar que o relacionamento entre as variáveis é fortemente linear?

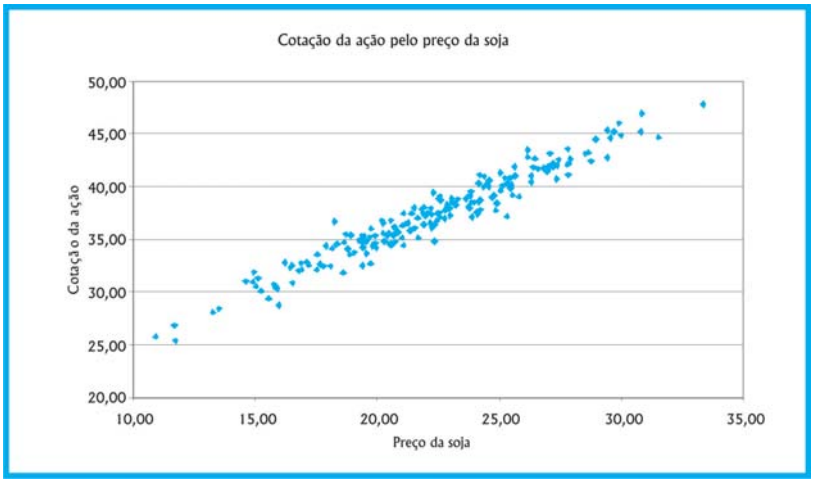


Figura 25: Diagrama de dispersão de Cotação da ação por Preço da soja
Fonte: adaptada pelo autor de Microsoft

GLOSSÁRIO

*Série temporal – conjunto de observações de uma variável quantitativa, ordenado no tempo (diário, semanal, mensal, anual). Fonte: Moore, McCabe, Duckworth e Sclove (2006).

A correlação entre as variáveis é claramente positiva: maiores valores de preço da soja correspondem a maiores valores de cotação da ação, o que parece plausível. A correlação parece ser muito forte, pois os pontos estão muito próximos. Quanto à natureza, pode-se observar que seria possível ajustar uma reta entre os pontos. Portanto, conclui-se que o relacionamento entre as variáveis é fortemente linear. Poderíamos, então, obter a equação da reta, para, a partir dos valores da soja, prever a cotação da empresa agroindustrial.

Se uma das variáveis quantitativas for o tempo (medido em anos, meses, semanas, dias, trimestres), teremos uma **série temporal***.

Em Saiba mais, vamos apresentar algumas referências sobre o assunto, que serão extremamente úteis, caso você tenha que lidar com séries temporais.

Você já deve ter visto em algum lugar uma tabela ou um gráfico mostrando a evolução do PIB do Brasil ao longo dos anos, ou a evolução da população de um país, ou mesmo os percentuais de intenção de voto dos candidatos a presidente em cada pesquisa eleitoral. O objetivo da análise de uma série temporal é identificar a existência de padrões que nos auxiliem a tomar decisões.

Vamos agora ao último caso desta Unidade, muito importante para o administrador, pois é bastante comum ele ter que estudar o relacionamento entre uma variável qualitativa e outra quantitativa.

Caso de uma variável qualitativa e uma quantitativa

Usualmente, pressupõe-se que analisaremos a variável quantitativa em função dos valores da variável qualitativa, visto que esta última costuma ter menos opções, o que simplificaria o processo e permitiria resumir mais os dados.

Na Unidade 1, falamos sobre classificação das variáveis por nível de manipulação em independente e dependente. Se estudamos duas variáveis, uma qualitativa e outra quantitativa, a qualitativa será considerada independente (ou de agrupamento), e a quantitativa, a dependente. Vejamos dois exemplos rápidos.

Imagine que você está realizando uma pesquisa experimental. Há interesse em avaliar a resposta a um medicamento contra o diabetes, que deveria reduzir o nível de glicose no sangue dos indivíduos portadores da doença. Para testar a eficiência do medicamento, você realiza um experimento, sorteando dois grupos de voluntários; um grupo receberá o medicamento, e o outro, o placebo durante um período de tempo. Ao final do experimento, os níveis de glicose dos indivíduos dos dois grupos são medidos para avaliar se no grupo que recebeu o medicamento eles sofreram redução significativa. Há duas variáveis, a independente, grupo de indivíduos, com dois valores (grupo tratado e grupo placebo), qualitativa, e a dependente, nível de glicose no sangue, quantitativa. Neste caso, a definição de variável independente como a que é manipulada para causar um efeito na dependente é aceitável.

Em outra situação, em uma pesquisa de levantamento, a variável independente seria meramente uma variável de agrupamento, para categorizar a variável dependente. Vamos ver um exemplo a respeito.

Para a mesma situação do Exemplo 1. Neste caso, gostaríamos de avaliar se existe algum relacionamento entre a renda do consumi-

dor e o modelo adquirido. Espera-se que exista tal relacionamento, pois os modelos Chiconaultla e DeltaForce3 são os mais baratos, e o sofisticado LuxuriousCar é o mais caro de todos.

Neste caso, podemos obter distribuições de freqüências da variável Renda para cada valor da variável Modelo. Seria uma situação semelhante à do item b do Exemplo 4, mas agora com cinco tabelas, uma para cada opção de Modelo.

Muito importante! Se optarmos por agrupamento em classes, todas as tabelas precisam ter o mesmo número de classes e as mesmas amplitudes de classe, para que possamos comparar os grupos. No nosso caso, vamos usar as classes obtidas no item b do Exemplo 4 para as cinco tabelas:

1,795|-18,645 18,645|-35,495 35,495|-52,345
52,345|-69,195 69,195|-86,045

Basta, então, ordenar as rendas em função dos modelos e contar as freqüências em cada modelo, resultando os dados ilustrados no Quadro 9:

RENDA	MODELO					Total
	Chiconautla	DeltaForce3	Valentiniana	SpaceShuttle	LuxuriousCar	
1,795|-18,645	73	20	4	0	0	97
18,645|-35,495	7	35	32	24	4	102
35,495|-52,345	1	1	4	18	14	38
52,345|-69,195	0	0	1	0	8	9
69,195|-86,045	0	0	0	0	3	3
Total	81	56	41	42	29	249

Quadro 9: Distribuições de freqüências de Renda agrupadas em classe por Modelo
Fonte: elaborado pelo autor

Há 249 dados no quadro, porque o dado perdido (descoberto no Quadro 7) foi removido do conjunto.

Observe a semelhança do quadro mostrado acima com o Quadro 7. Da mesma forma que lá fizemos, é possível calcular percentuais em relação aos totais das linhas, colunas ou total geral.

Podemos perceber que o relacionamento esperado entre as variáveis foi confirmado: para os modelos mais baratos, a renda mais alta

está na classe de 35,495 a 52,345 salários mínimos; já os clientes do modelo mais caro (LuxuriousCar) estão nas classes mais altas.

O Quadro 9 poderia ser expresso através de um gráfico, um **histograma categorizado**. Infelizmente, tal gráfico não pode ser feito em uma planilha eletrônica (como o Excel®) sem consideráveis manipulações. Mas, através de um software estatístico, no nosso caso, o Statsoft Statistica 6.0®, isso é possível (Figura 26):

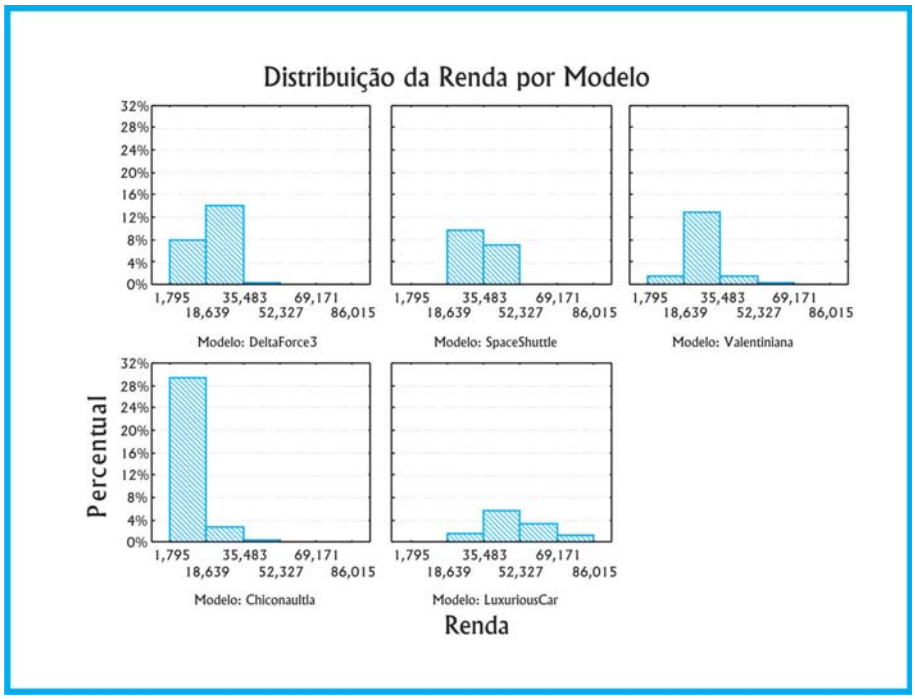


Figura 26: Histograma categorizado da Renda por Modelo adquirido
Fonte: adaptada pelo autor de Statsoft®

Observe que o software dividiu a variável Renda em cinco classes também, mas com limites ligeiramente diferentes dos nossos. Além disso, optamos por apresentar os resultados em percentuais relativos ao total dos dados (249). A interpretação é semelhante à da tabela.

Na prática, o mais comum, quando analisamos uma variável quantitativa em função de uma qualitativa, é calcular medidas de síntese daquela para cada grupo definido pelos valores desta. A partir dos resultados, é possível verificar se existe relacionamento entre as variáveis. Veremos na Unidade 4 as medidas de síntese.

Saiba mais...

- Sobre correlação entre variáveis (quantitativas): BARBETTA, P. A. *Estatística Aplicada às Ciências Sociais*. 6. ed. Florianópolis: Ed. da UFSC, 2006, capítulo 13.

- Sobre correlação entre variáveis (quantitativas): MOORE, D. S.; et al. *A prática da Estatística empresarial*: como usar dados para tomar decisões. Rio de Janeiro: LTC, 2006, capítulo 2.

- Sobre análise de séries temporais: LEVINE, D. M.; et al. *Estatística*: teoria e aplicações – usando Microsoft Excel em português. 5. ed. Rio de Janeiro: LTC, 2005, capítulo 13.

- Sobre análise de séries temporais: STEVENSON, Willian J. *Estatística Aplicada à Administração*. São Paulo: Harbra, 2001, capítulo 16.

- Sobre como realizar as análises descritas nesta Unidade e na Unidade 4 através do Microsoft Excel®, consulte "Como realizar análise exploratória de dados no Microsoft Excel®", disponível no Ambiente Virtual de Ensino-Aprendizagem, assim como o arquivo de dados usado nos exemplos apresentados.

RESUMO

O resumo desta Unidade está mostrado na Figura 27:

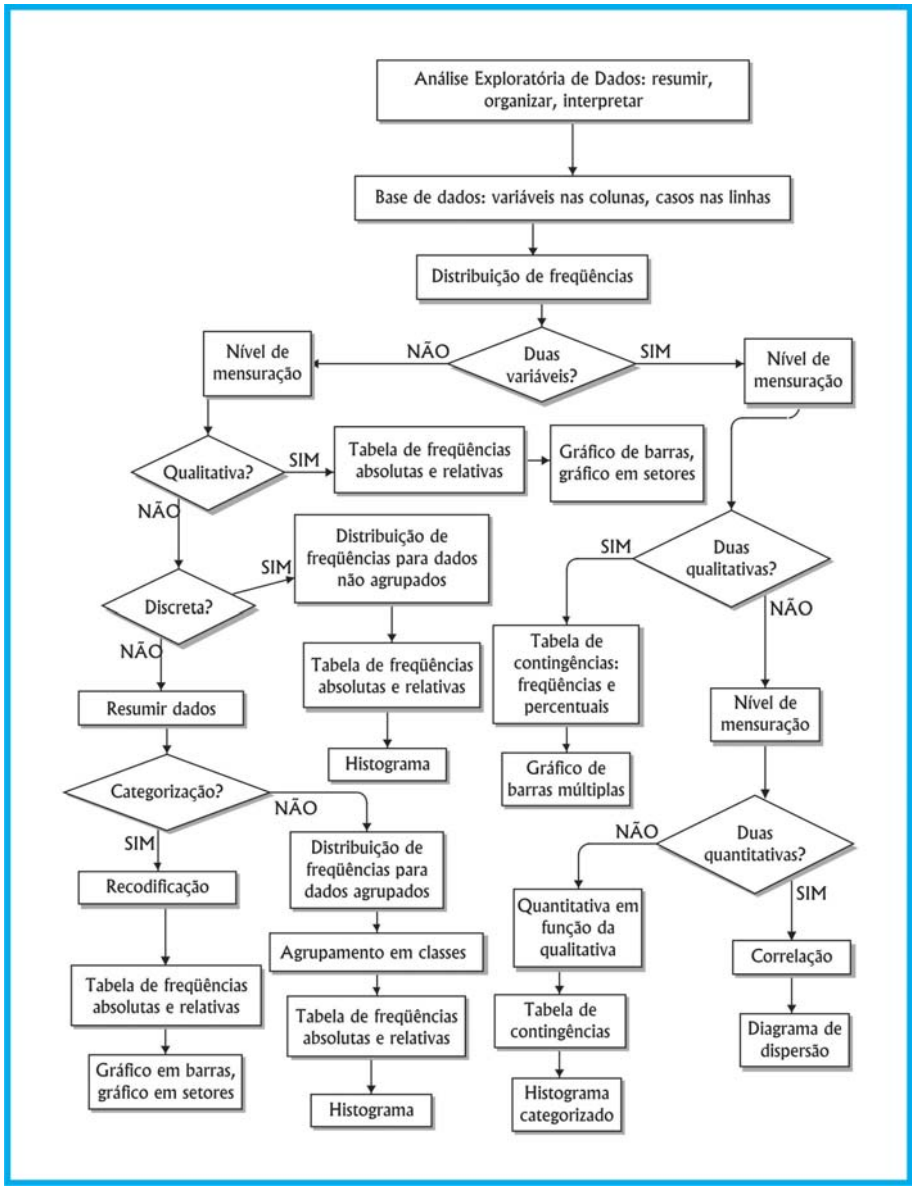


Figura 27: Resumo da Unidade 3
Fonte: elaborada pelo autor

Atividades de aprendizagem

As atividades de aprendizagem estão disponíveis no Ambiente Virtual de Ensino-Aprendizagem. Não deixe de respondê-las.

Caro estudante!

Esta Unidade foi importantíssima para você entender a Análise Exploratória de Dados. Vimos como organizar, interpretar e resumir as informações coletadas, os níveis de mensuração e o número de variáveis. Você aprendeu a elaborar tabelas, planilhas e gráficos de acordo com as especificidades das informações colhidas. Chegamos ao final da Unidade e ao começo de uma nova aprendizagem. Esta Unidade lhe deu base para o aprendizado proposto nas Unidades seguintes. Leia e releia quantas vezes sejam necessárias os variados exemplos propostos para cada categoria estudada. As figuras, os quadros, as representações e os exemplos são grandes aliados nesse processo de aprendizagem.

Interaja com sua turma e responda as atividades. A tutoria está pronta a lhe auxiliar, e o professor, ansioso em reconhecer suas habilidades desenvolvidas a partir do conhecimento deste conteúdo. Vamos em frente!