COPPEAD/UFRJ

RELAÓRIO COPPEAD Nº 181

OS DADOS E SUA NATUREZA

Eduardo Saliby *

Fevereiro de 1987

* Professor adjunto e pesquisador da COPPEAD-Instituto de Pós-Graduação e Pesquisa em Administração/UFRJ.

OS DADOS E SUA NATUREZA

Eduardo Saliby

Fevereiro de 1987

APRESENTACÃO

Muitas das dificuldades encontradas no aprendizado e uso da estatística decorrem, a nosso ver, de uma lacuna existente entre a forma como ela é normalmente ensinada e a forma com que suas aplicações emergem em trabalhos práticos ou de pesquisa.

Por exemplo, a matéria prima de qualquer análise estatística são os dados, a partir dos quais queremos tirar algum tipo de conclusão. Ora, na prática são raros os casos em que estes dados satisfazem plenamente todos os requisitos exigidos para o uso de uma determinada técnica estatística. Por exemplo, enquanto a maioria dos textos nos ensinam a analisar dados oriundos de uma amostragem aleatória, uma situação prática muito comum é aquela em que não temos nenhum controle sobre a sua coleta. Tais dados, apesar de não terem uma origem aleatória, certamente contém valiosas informações que merecem ser consideradas.

Assim sendo, identificamos um desbalanceamento entre a atenção dedicada às ferramentas estatísticas e aquela dada aos próprios dados. Ora, todos sabemos que a qualidade final de um produto depende igualmente tanto da sua matéria prima como do seu processamento. Não custa enfatizar que, mesmo com a técnica mais sofisticada e o computador mais avançado, jamais conseguiremos chegar a conclusões válidas, se tivermos como ponto de partida dados de má qualidade.

O presente texto estuda a natureza dos dados, sem maiores preocupações com o ferramental estatístico disponível para sua análise.

Os Dados e sua Natureza

"Não são os dados que mentem,
somente aqueles que os interpretam."

"Os números governam o mundo."

1. Introdução: O Que São e Para Que Servem os Dados?

Vivemos hoje na era da informação e os dados são também um tipo de informação. Mas o que são e para que servem os dados?

Os dados resultam da observação da realidade, através de um procedimento de mensuração, contagem ou classificação. Sem maiores rigores formais, podemos dizer que dados são informações, não necessariamente numéricas, organizadas num formato específico, que descrevem características quantitativas ou qualitativas de um grupo de elementos. Os dados são essenciais tanto na investigação científica como na tomada de decisões.

Embora cada situação tenha sua motivação específica, podemos dizer que os dados servem tanto para prover novos conhecimentos, sugerindo assim novas teorias, como também para a verificação empírica destas teorias. É também com base em dados que as decisões são tomadas, avaliadas e, se for o caso, modificadas.

No presente trabalho, consideraremos os dados unicamente como elementos do processo de generalização. Neste contexto, um dado único e individual seria de pouca ou nenhuma utilidade pois qualquer generalização feita a partir dele se transformaria num "tiro no escuro". Toda generalização é sujeita a erros, mas para que se tenha maiores chances de sucesso, os dados em que se baseia devem ser objetivos e confiáveis.

Dados objetivos são aqueles que, em teoria, independem do observador, sendo portanto reprodutíveis. Tal situação representa antes um ideal a seguir do que uma regra, uma vez que na prática, são raras as ocasiões em que se tem uma total reprodutibilidade dos dados. Por isso, a ciência se mostra tolerante para com pequenas distorções nos dados, desde que decorrentes de fatores fora do controle do operador.

Já a alteração intencional dos dados, fornecendo um quadro distorcido da realidade, é uma prática altamente condenável na ciência (1); tais fraudes, quando descobertas, levam muitas vezes o seu autor a cair em desgraça nos meios científicos. Pena que não se faça o mesmo com os políticos...

---

(1) A este respeito, recomendamos a leitura do artigo de Vieira, S. Fraude em Ciência. Ciência Hoje, 5 (25): 74-79, jul./ago. 1986.

A confiabilidade diz respeito à capacidade dos dados realmente refletirem as características que desejamos estudar. De nada adianta dispormos de uma enorme quantidade de dados, das melhores técnicas estatísticas, do melhor computador e da melhor inteligência, se eles não descreverem fielmente a realidade.

A qualidade dos dados é, em última instância, determinada pelo processo de coleta. A menos que tenhamos informações suplementares sobre o fenômeno em estudo, o que felizmente muitas vezes ocorre, nunca teremos condição de avaliar a qualidade de um conjunto de dados, pela sua simples inspeção. Assim sendo, só teremos condição de saber se os dados de idade para um grupo de alunos são corretos ou não, se dispusermos de algum tipo de informação adicional a este respeito, como por exemplo a sua faixa esperada de variação.

Por isso, ao estudarmos um conjunto de dados, ainda que com objetivos meramente descritivos, é sempre importante sabermos qual foi o processo que lhes deu origem. Conforme veremos adiante, o processo de geração também determina a escala ou nível de mensuração de cada variável, cuja identificação representa um dos passos iniciais em qualquer estudo. Somente após conhecermos a natureza dos dados, identificando suas principais características, é que teremos condição de escolher as ferramentas estatísticas mais apropriadas para sua análise e interpretação.

2. Dados Individuais e Dados Consolidados

A coleta de dados sempre resultará num conjunto de um ou mais valores para cada elemento observado. Cada valor descreve uma das características em estudo, como por exemplo a idade, o peso e o estado civil de um aluno.

Concluída a fase de coleta, teremos um conjunto de "dados brutos". Antes, porém, de passarmos à sua análise, deveremos nos preocupar com sua qualidade. Assim sendo, o próximo passo do processo é, geralmente, a etapa conhecida como crítica dos dados. Procuraremos identificar eventuais problemas que os dados venham a apresentar, como erros de coleta, transcrição ou mesmo casos em que uma observação venha a se mostrar totalmente atípica.

Dependendo do problema apresentado e do estudo em questão, os dados que não passarem pela crítica poderão ser corrigidos ou então eliminados. Com isso, chega-se a um conjunto de dados individuais, prontos para análise; usando-se computador, estes dados seriam armazenados num arquivo ou banco de dados.

Como a estatística não se preocupa com o estudo de dados individuais, mas somente de conjuntos de dados, qualquer análise estatística resultará na sua consolidação. Esta consolidação, que nada mais é do que uma descrição mais suscinta, poderá ser feita através de:

- números, como sua média e desvio padrão,
- tabelas, ou
- gráficos.

Embora resulte na perda da informação individual sobre cada elemento, a consolidação dos dados é indispensável para sua interpretação. É também na forma consolidada que os dados são geralralmente comunicados a terceiros, como aqueles fornecidos pelo IBGE ou aqueles que apresentamos numa reunião ou relatório.

A consolidação dos dados não requer, normalmente, o uso de procedimentos matemáticos complexos. Mas, isto não quer dizer que ela tenha menor importância em relação às demais ferramentas estatísticas. Tanto assim que, uma das partes em que a Estatística se divide, a Estatística Descritiva estuda somente os métodos de consolidação dos dados.

Mas, para que esta consolidação possa ser feita de maneira mais sistemática, os dados individuais deverão ser dispostos num formato apropriado para o seu processamento. Isto será visto a seguir...

3. A Organização dos Dados Individuais

Em princípio, todo estudo requer um diferente conjunto de dados, específico para cada situação. Há casos em que estes dados não estão disponíveis, e, por isso, deverão ser coletados. Há casos, porém, em que os dados já foram previamente coletados e se encontram disponíveis, como por exemplo num banco de dados; ainda assim, somos normalmente obrigados a reorganizá-los.

Hoje em dia, em função da maior disponibilidade de recursos de hardware, software e transmissão de dados, a sua preparação para análise é tarefa das mais simples. Um conjunto ou arquivo de dados, é comumente organizado segundo uma matriz.

Esta matriz terá uma linha para cada elemento observado. Cada linha contém as informações de interesse sobre cada elemento, definindo assim uma observação ou caso.

A matriz terá também uma coluna para cada característica em estudo, cada uma correspondendo a uma diferente variável.

Um exemplo nos ajudará a melhor compreender a estrutura desta matriz:

Consideremos um conjunto de dados bem simples, compreendendo o nome, a idade, a altura e o peso de 5 crianças:

Nome	Idade (anos)	Altura (cm)	Peso (Kg)
Juliana	11	148	31
Carolina	7	125	25
Paulo	10	146	35
Mariana	10	142	31
Alexandre	5	111	23

Neste caso, temos N = 5 observações, uma para cada criança, e M = 4 variáveis: Nome, Idade, Altura e Peso. As 3 últimas são do tipo numérico, enquanto o Nome é uma variável qualitativa.

Embora um trabalho prático envolva normalmente um maior número de observações e de variáveis, os dados poderão ser sempre dispostos segundo uma matriz. Assim por exemplo, um questionário com um total de 20 perguntas, aplicado a um grupo de 100 pessoas, resultará numa matriz com M = 20 colunas (uma para cada pergunta), e N = 100 linhas (uma para cada entrevistado).

Este tipo de organização dos dados é utilizada pela maioria dos programas para análise estatística, como por exemplo o SPSS (Statistical Package for the Social Sciences), o SAS (Statistical Analysis System) e o Statgraphics.

4. A variabilidade estatística

Não havendo variabilidade num conjunto de dados, não existiria motivo para uso da estatística. Imagine por um instante um mundo (nem um pouco admirável) em que tudo acontecesse da mesma maneira, onde todas as coisas fossem iguais, assim como as pessoas, suas opiniões e seus gostos. Num mundo assim, aliás muito sem graça, o estatístico morreria de fome!

Felizmente para nós, e em particular para os estatísticos, a natureza é pródiga em variações, apesar de preservar de forma sábia uma ordem ou padrão nestas variações. Não temos nenhum interesse no estudo de características com pouca ou nenhuma variação, como por exemplo o número de cozinhas em apartamentos tipo "classe média". A própria ciência tem por objetivo o estudo das variações, procurando explicá-las através de teorias; aliás, esta é a razão que nos leva a utilizar o termo variável para denotar uma característica de interesse.

Sendo os dados um registro da realidade, sua variabilidade poderá ter dupla origem:

a) No próprio elemento observado; neste caso, ela é denominada variabilidade interna ou intrínseca.

b) No processo de observação, também denominada variabilidade externa ou extrínseca.

A variabilidade interna é aquela inerente ao elemento observado. Um exemplo deste tipo de variabilidade seria o número de filhos de um casal ou o saldo de depósitos de uma agência bancária.

Já a variabilidade externa é aquela introduzida durante o processo de observação. Nas situações mais simples, ela pode ser atribuída às limitações do instrumento de medida ou a erros de seu operador. Em situações mais complexas, ela reside na dificuldade de operacionalização, ou seja, na dificuldade de definição de um procedimento apropriado de medida. Um bom exemplo do blema de operacionalização é a mensuração da taxa de inflação, para a qual dispomos de vários índices, cada um levando a diferentes valores, ainda que para um mesmo período de tempo.

Embora com origem distinta, ambas as variações - interna e externa - recebem um mesmo tratamento estatístico. No entanto, sem que com isto subestimemos a importância da variabilidade externa e de suas possíveis conseqüências nossa atenção volta-se normalmente ao estudo da variabilidade interna. Para explicar esta variabilidade, adota-se, geralmente, uma abordagem baseada na sua decomposição em duas parcelas: uma determinística e a outra aleatória ou probabilística.

Variabilidade determinística:

A variabilidade determinística é aquela que pode ser totalmente explicada através de relações com outras variáveis. Dependendo do nível de precisão desejado, uma interpretação determinística poderá ser ou não apropriada, apesar de serem raras as situações em que se tenha um comportamento essencialmente determinístico. Em ciências sociais, por exemplo, nunca conseguimos um bom nível de explicação usando modelos determinísticos!

Exemplos de situações em que uma descrição determinística é adequada, são as chamadas leis físicas (o que justifica o uso da palavra "lei"), como por exemplo:

a) a lei da gravidade;
b) a lei de Ohm;
c) as leis da termodinâmica.

Variabilidade aleatória:

Caracteriza-se por ser, em princípio, imprevisível, mas passível de uma descrição probabilística. Exemplos deste tipo de variação seriam o consumo de energia elétrica de uma residência, o número de acertadores da quina num concurso da loto, ou o intervalo de tempo entre emissões sucessivas de partícula alfa por um material radioativo.

Quanto a sua origem, esta variação pode ter uma das seguintes interpretações:

a) Ela é o resultado de causas ainda desconhecidas ou pouco conhecidas.

b) Ela é o resultado de múltiplas causas que, embora pudessem ser identificadas, não teríamos nenhuma vantagem em fazê-lo.

c) A terceira e última interpretação da variabilidade aleatória é motivo de muita discussão filosófica.
Assim, caso você concorde, esta variação poderia ser vista como uma característica inerente à própria natureza, a qual teria também um comportamento aleatório.

A este respeito é interessante mencionar alguns trechos da calorosa discussão havida entre Einstein (um determinista) e Bohr (um não-determinista). Criticando a interpretação não-determinista, Einstein afirmou que "Deus não joga dados com o mundo!" Bohr, em defesa de seu ponto de vista retrucou a Einstein: "Não diga a Deus o que Ele pode ou não fazer!"

Qualquer que seja a interpretação da variabilidade aleatória, a sua descrição através de modelos probabilísticos é sempre válida. É interessante notar que, apesar de descreverem variações aleatórias, estes modelos são definidos através de parâmetros supostos constantes ou invariantes. Por exemplo, quando afirmamos que a probabilidade de se tirar cara com uma moeda é 1/2, estamos fixando os parâmetros da distribuição de probabilidades associada a este experimento. Seguindo a mesma linha de raciocínio, vemos também que os parâmetros relativos a dados populacionais são também constantes, ainda que muitas vezes desconhecidos.

A prática científica consiste na identificação e na decomposição das variações observadas nestas duas componentes: a determinística e a aleatória. À medida em que aumenta a contribuição da parcela determinística, aumenta também o nosso poder de previsão sobre os resultados futuros. Cabe mencionar no entanto, que a distinção entre os dois tipos de variação é um problema complexo, que implica na busca de relações de causa e efeito. Uma dificuldade adicional a ser citada é que, em função de nosso conhecimento ser sempre incompleto, nunca teremos certeza de que uma variação hoje atribuída a causas aleatórias não será algum dia explicada através de relações determinísticas.

5. A classificação dos dados estatísticos

Os principais critérios para a classificação dos dados estatísticos são os seguintes:

a) Quanto a sua fonte.
b) Quanto ao método de coleta.
c) Quanto a sua abrangência.
d) Quanto ao número de variáveis ou dimensões.
e) Quanto a dimensão temporal.
f) Quanto ao nível de agregação.
g) Quanto ao nível ou escala de mensuração.

Embora existam outras maneiras para se classificar um conjunto de dados, os critérios acima são os mais relevantes na escolha dos procedimentos estatísticos para sua análise.

A seguir, vamos estudar cada um destes critérios com maior detalhe.

Classificação quanto a sua fonte:

Uma das decisões iniciais em qualquer estudo é quanto as possíveis fontes de dados. Embora a coleta direta dos dados seja aparentemente o caminho mais simples, ela deve se restringir aos casos em que não dispomos de outra alternativa a seguir. O uso de dados previamente coletados irá nos poupar de uma tarefa geralmente trabalhosa, demorada e cara. Alguns autores chegam ao ponto de estabelecer um paralelo entre a coleta direta de dados e a cirurgia clínica, ambos para serem utilizados quando as demais alternativas de ação mostrarem-se inviáveis.

Dependendo de sua fonte, mas não do conteúdo, poderemos classificar os dados em primários e secundários.

Dados primários são aqueles gerados para atender as necessidades específicas de um particular estudo. Este é o caso quando saímos a campo para sua coleta. Exemplos de dados primários são aqueles provenientes da aplicação de questionários e entrevistas, as notas de alunos numa prova, ou a duração de uma operação produtiva.

Dados secundários são aqueles previamente disponíveis ou fornecidos por terceiros, coletados sem o objetivo expresso de atender as necessidades específicas do estudo em questão. Poderíamos também chamá-los de "dados de segunda-mão", sem que tenham, porém, a conotação de um material de pior qualidade ou obsoleto; ao contrário, quando disponíveis, os dados secundários tendem a ser de melhor qualidade!

Dependendo da localização da sua fonte em relação ao responsável pelo estudo, os dados secundários podem ser ainda in-

ternos ou externos.

Dados internos são aqueles gerados dentro de uma empresa ou instituição, porém utilizados com fins diferentes daqueles que lhes deram origem. Este é o caso, por exemplo, quando os dados de notas fiscais, gerados pelo faturamento de uma empresa, são utilizados pela área de "marketing" em análises de vendas por produto ou por região.

Os dados externos são geralmente fornecidos por instituições credenciadas como o IBGE, a Fundação Getulio Vargas, a DATAPREV ou empresas especializadas neste tipo de serviço. Exemplos de dados secundários são os indicadores sócio-econômicos, dados demográficos, arrecadação de impostos, consumo total de energia e estatísticas de produção industrial. Atualmente, muitos destes dados podem ser diretamente acessados através de telex, videotexto, terminais de computador ou microcomputadores. Este é o caso, por exemplo, dos sistemas SIDRA do IBGE e SINTESE da DATAPREV.

Apesar da utilidade e importância dos dados secundários num estudo sócio-econômico ou num trabalho de pesquisa, a atenção da estatística volta-se, primordialmente, para o estudo de dados primários. Justifica-se tal abordagem por dois motivos:

Em primeiro lugar, os dados secundários são geralmente fornecidos já prontos para uso. Assim, em princípio, só deveremos nos preocupar com a preparação e a consolidação de dados primários.

O segundo e principal motivo para uma maior atenção aos dados primários são as inferências. Neste caso, o processo de geração dos dados é um fator determinante das generalizações que poderemos fazer a partir deles. Sendo assim, somos normalmente obrigados a coletá-los, sumariá-los, interpretá-los e analisá-los, usando para isso o ferramental estatístico disponível.

Classificação quanto ao método de coleta:

Os dois métodos gerais de coleta de dados são a observação e a experimentação. Enquanto a observação procura estudar um fenômeno como ele ocorre na vida real, a experimentação caracteriza-se por uma manipulação controlada da situação em estudo. Assim sendo, a observação resulta numa menor interferência do observador sobre o fenômeno, mas tem o inconveniente de possibilitar um menor controle sobre os elementos selecionados e/ou condições de observação. Com a experimentação dá-se justamente o contrário.

Esta diferenca entre os dois métodos de coleta refere-se, pois, ao grau de interferência do observador no fenômeno. Na prática de pesquisa, são raros os estudos observacionais totalmente livres da influência do observador; uma situação em que se supõe não haver este tipo de interferência, são as observações astronômicas.

Mas, em ciências sociais, a situação é outra. Aqui, além do problema da seleção dos elementos a serem observados, geralmente resolvido através de uma escolha aleatória, outro importante problema decorre do fato de que o ato de observar, por mais cuidadoso que seja, sempre terá algum tipo de interferência sobre o indivíduo. Dá-se, a este fenômeno, o nome de "efeito do pesquisador".

Uma forma, às vezes bem sucedida, para se controlar este efeito consiste em se fazer com que um indivíduo ignore estar sendo observado. Neste caso, existem importantes questões éticas a serem consideradas, uma vez que esta abordagem pode resultar num verdadeiro trabalho de espionagem.

Já a experimentação em ciências sociais, tem o inconveniente de introduzir um controle excessivo sobre as condições de observação, tornando mais arriscada uma generalização de seus resultados. Uma solução de compromisso seria o uso de experimentos simulados onde as condições de observação, ou mesmo os elementos observados, seriam apenas fictícios. Com este recurso, poderíamos artificialmente reproduzir situações mais complexas, e mais proximas da realidade.

Classificação quanto a sua abrangência:

Quanto a abrangência, podemos classificar os dados em populacionais e amostrais. Esta classificação, apesar de irrelevante quanto ao aspecto descritivo, é conceitualmente bastante importante. Isto porque, enquanto os dados populacionais tem finalidade unicamente descritiva, os dados amostrais tem, geralmente, finalidades inferenciais.

Embora não seja correto, em princípio, fazer inferências a partir de dados populacionais, cabe lembrar que a distinção entre amostra e população é arbitrária, sendo função do particular estudo. Assim sendo, poderemos ver uma população como um caso particular de amostra, que dentre todas as possíveis realizações da natureza, se materializou. Segundo esta interpretação, uma população passa a ser vista como uma observação da macro-população definida por todas as realizações possíveis; como tal, ela é passível de inferências.

De qualquer forma, sempre que se fizer inferências a partir de dados populacionais, esta hipótese de trabalho deverá ser claramente indicada.

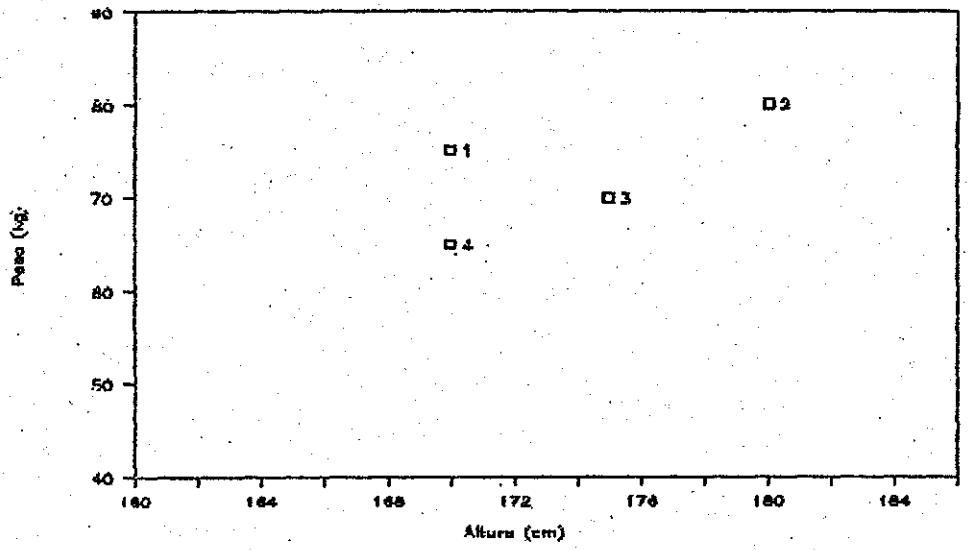
Classificação quanto ao número de variáveis ou dimensões:

O número de variáveis corresponde ao número de colunas da matriz de dados. Considere, por exemplo, os dados de altura e peso de um conjunto de 4 alunos, a saber:

Aluno	Altura (cm)	Peso (Kg)
1	170	75
2	180	80
3	175	70
4	170	65

Definindo a altura e o peso dos alunos como eixos coordenados, poderemos representar cada aluno por um ponto. Isto é mostrado na figura 1, onde cada aluno é identificado pelo respectivo número.

Figura 1 : Alturas e pesos de 4 alunos



Convém observar que caso tivéssemos dois ou mais alunos com exatamente as mesmas medidas de peso e altura, seus pontos iriam coincidir.

Podemos facilmente generalizar esta idéia para mais de duas dimensões. Por exemplo, se além da altura e do peso, considerássemos também a idade de cada aluno, uma observação seria agora definida por um ponto num espaço tri-dimensional; neste caso, o terceiro eixo é definido pela variável idade.

Vemos, assim, que cada variável corresponde a uma diferente dimensão do espaço de dados. Na situação mais geral em que tivermos M variáveis, poderemos imaginar cada observação como um ponto num espaço M-dimensional. Embora só possamos visualizar espaços com até 3 dimensões, esta interpretação é um recurso que pode nos ajudar na análise de dados.

É claro que a situação mais simples é aquela em que temos uma única variável, ou seja, quando nossos dados são univariados. Neste caso, cada observação é definida por um único valor, correspondendo a um caso especial em que a matriz de dados se reduz a uma só coluna. Consequentemente, utilizaríamos apenas um eixo de coordenadas para sua representação geométrica.

Muitas técnicas estatísticas destinam-se ao estudo de dados univariados e, por isso, são também chamadas técnicas univariadas. Exemplos destas técnicas são o cálculo de médias e todos os testes estatísticos com ela relacionados.

A situação mais geral, no entanto, é aquela em que temos duas ou mais variáveis. Neste caso, nossos dados são do tipo multivariado.

É importante notar que, mesmo trabalhando com dados multivariados, poderemos estudar as variáveis em separado, dando-lhes, pois, um tratamento univariado. De fato, este é o procedimento usual pelo qual iniciamos um estudo, sejam os dados univariados ou multivariados. Análises mais simples, onde procuramos entender o comportamento individual de cada variável, devem preceder estudos mais complexos onde procuramos relacioná-las.

A busca de relações entre variáveis, estudando sua variação conjunta, envolvem o uso de métodos estatísticos mais sofisticados: as técnicas multivariadas. Algumas destas técnicas são bem conhecidas, como a análise de regressão e correlação; outras, como a análise fatorial, constituem tópicos de estudo mais avançado.

Sem entrar em maiores detalhes quanto as técnicas multivariadas, podemos classificá-las em métodos de análise de dependência ou de interdependência. Nos métodos de dependência, procuramos prever uma ou mais variáveis, ditas dependentes, a partir de um conjunto de variáveis explicativas, ou independentes. Este é, por exemplo, o caso da análise de regressão.

Nos métodos de interdependência, não existe uma preocupação em se prever ou explicar o comportamento de uma variável em função das demais, mas apenas estudar o seu comportamento conjunto. Exemplos deste tipo de abordagem são a análise de correlação e a análise fatorial.

Não há, em princípio, um limite teórico a ser estabelecido quanto ao número de variáveis a serem consideradas num particular estudo. No entanto, razões de ordem prática nos impõem restrições.

Em primeiro lugar, cabe lembrar que quanto maior for o número de variáveis, mais complexas e demoradas se tornam a coleta e a preparação destes dados. Outra razão, e a nosso ver a mais importante, é que a própria análise se torna mais difícil na esperança de um maior número de variáveis. Isto ocorre não só porque as possibilidades de combinacões entre as variáveis cresce enormemente, como também porque a interpretação de relacões envolvendo muitas variáveis é tarefa das mais complexas.

Classificação quanto a dimensão temporal:

Muitas vezes, notadamente em estudos econômico-financeiros, nossos dados constituem uma sequência de observações ordenadas no tempo, denominada série temporal ou histórica. Um exemplo de série histórica é a produção anual de petróleo bruto pelo Brasil:

Produção de Petróleo Bruto - Brasil (1975 a 1986)

Ano	Produção (Milhões de metros cúbicos)
1975	9 979
1976	9 702
1977	9 332
1978	9 304
1979	9 608
1980	10 562
1981	12 384
1982	15 082
1983	19 140
1984	26 837
1985	31 724
1986	33 201

Fonte: PETROBRÁS.

Além de ordenadas no tempo, as observações de uma série referem-se a períodos de idêntica duração ou a datas igualmente espaçadas. A este intervalo de tempo, que no nosso exemplo é anual, denominamos periodicidade da série.

As séries históricas poderão ser também do tipo univariado ou multivariado. No caso de séries multivariadas, elas deverão ter sempre a mesma data de origem, duração e periodicidade.

Outra classificação das séries, que diz respeito à natureza do fenômeno descrito, é em serie de estoques ou de fluxos.

Uma série é de estoques quando seus valores, descrevendo posições, podem ser vistos como uma sequência de fotos tomadas a intervalos igualmente espaçados no tempo. A título de exemplo de série de estoques, poderíamos citar: o volume de depósitos do FGTS, o montante das reservas cambiais do país, o preço diário de uma ação, o valor mensal do salário mínimo ou a cotação diária do dólar.

Já uma série é dita de fluxos quando seus valores descrevem variações observadas entre suas datas de referência. Exemplos de séries de fluxos são as estatísticas de produção e consumo, de entrada e saída de capital, de nascimentos e mortes.

Ao contrário das séries históricas, existem também dados em que a dimensão tempo não é levada em conta. Uma destas situações

ocorre nos chamados estudos longitudinais, em que todas as observações referem-se a um único momento. Outra situação é quando a dimensão tempo é ignorada por ser considerada irrelevante; este é o caso, por exemplo, quando comparamos valores de Q.I. para diferentes grupos de alunos.

Os dados não-temporais são, em geral, desprovidos de ordenação. Isto significa que, ao contrário dos dados temporais, poderemos alterar livremente a ordem de suas observações.

Classificação quanto ao nível de agregação:

Em alguns casos os dados poderão ser agregados em vários níveis, caracterizando uma ordem hierárquica. Os critérios mais usuais para esta agregação são de natureza geográfica ou temporal. Por exemplo, algumas estatísticas fornecidas pelo IBGE através do sistema SIDRA, tais como o grau de instrução da população apurado no censo demográfico, podem ser agregadas nos seguintes níveis:

Município,
Microrregião
Mesorregião
Região Metropolitana
Unidade da Federação
Região Geográfica
Brasil

Situacões similares, em que também agregamos os dados hierarquicamente, ocorrem em análises de vendas de produtos de empresas que trabalham com várias linhas e que atuam em diversos segmentos do mercado e em diferentes regiões.

Classificação quanto ao nível de mensuração:

Embora os dados estatísticos sejam normalmente apresentados sob a forma numérica, o significado destes números poderá variar substancialmente dependendo do processo que lhes deu origem: medida, contagem ou classificação. Uma exigência básica de qualquer processo de mensuração é que deverá sempre resultar na atribuição de um único valor para cada elemento observado.

O nível de mensuração é um atributo associado a cada uma das variáveis em estudo, de acordo com o seu significado numérico. É o nível de mensuração que determina os procedimentos estatísticos aplicáveis a cada variável. Por este motivo, sua identificação normalmente representa um dos passos iniciais, e dos mais importantes, em qualquer estudo.

Os quatro níveis de mensuração que poderemos encontrar são:

O nominal ou categórico,
O ordinal ou por postos,
O intervalar, e
O de razão.

A ordem com que estes quatro níveis de mensuração são apresentados corresponde ao grau crescente de informação que cada escala traduz. Assim sendo, uma variável expressa num nível intervalar nos dirá mais do que outra expressa num nível ordinal ou nominal.

Além disso, as propriedades válidas para um nível de mensuração mais baixo são também válidas para os níveis que lhe são superiores. Por exemplo, os dados em escala intervalar satisfazem todas as propriedades dos dados ordinais. Conseqüentemente, os métodos estatísticos que se aplicam aos níveis mais baixos de mensuração, também se aplicam aos níveis mais altos. Por exemplo, todas as análises feitas com dados ordinais poderão ser também feitas com dados intervalares, mas a recíproca não vale.

6. Níveis ou escalas de mensuração

Tendo em vista a importância da escala de mensuração na escolha do procedimento estatístico a ser empregado na análise de um conjunto de dados, iremos estudá-las com maior detalhe.

Nível Nominal:

Corresponde ao mais baixo nível de mensuração em que as observaçoëes são unicamente classificadas em grupos ou classes. Esta classificação é feita em função das similaridades apresentadas, de modo a se ter a máxima homogeneidade possível dentro de cada categoria. Exemplos deste tipo de escala são as classificações de empresa por setor de atividade, pessoas quanto ao estado de origem, carros quanto a cor, etc.

A classificação é o princípio de qualquer ciência. Qualquer estudo pressupoe uma classificação prévia dos elementos para que possam ser devidamente analisados. Para se ter uma idéia da importância do processo de classificação na ciência, basta lembrarmos da existência de uma área especifica de estudos, a taxonomia, que trata deste problema.

Embora possamos ter vários níveis hierárquicos de classificação, consideraremos aqui somente o caso mais simples em que temos um único nível. Uma das características da escala nominal é que, além de distintas, não existe qualquer outro tipo de relação entre as categorias definidas, como por exemplo uma ordenação por um critério de preferência.

O único requisito para uso de uma escala nominal é que ela resulte numa classificação dos elementos em categorias mutuamente exclusivas e exaustivas, ou seja, cada elemento será classificado em uma - e em somente uma - categoria.

Na prática, esta restrição poderá trazer-nos alguns problemas, pois não é sempre que uma classificação se aplica. Assim, para nos prevenirmos quanto a este problema numa pesquisa de opinião, deveríamos sempre prever uma opção do tipo "não se aplica" para cada pergunta. Por exemplo, numa pergunta em que solicitamos a um respondente que escolha sua marca preferida de cerveja, dentre uma lista de marcas apresentadas, deveremos prever a possibilidade do entrevistado não gostar desta bebida.

Quando dispomos de um computador para a apuração de nossos dados, é comum codificarmos as diferentes categorias de uma variável nominal atraves de números inteiros. Tomemos novamente nossa pesquisa de preferência de cerveja; teríamos algo assim:

(1) Cerveja A
(2) Cerveja B
(3) Cerveja C
(4) Nenhuma das anteriores

Sendo os dados expressos na escala nominal, operações aritméticas, como sua soma ou subtração não são, naturalmente, válidas. Neste caso, os números não passam de meros códigos utilizados em lugar dos nomes, não havendo sentido algum em, digamos, somar a cerveja A com a C, a menos que sua mistura esteja fazendo algum efeito sobre nós!

Embora seja claro neste exemplo que não podemos fazer cálculos com dados nominais, há situações em que isto não é tão evidente. Uma boa razão de alerta reside no fato de que a maioria dos programas de computador para análise estatística ignora a escala de mensuração dos dados, calculando muitas vezes medidas que não se aplicam à escala em uso.

Em termos de análise estatística, o máximo que podemos fazer com dados nominais é a contagem de frequências, fornecendo as medidas e procedendo aos testes com ela relacionados.

Nível Ordinal:

Esta escala, muito utilizada na área de ciências sociais, caracteriza-se por ser uma classificação em que as categorias dispõem, agora, de uma relação de ordem. Um exemplo deste tipo de escala é quando classificamos as empresas, em função de seu tamanho, em pequenas, médias ou grandes.

Um dos principais usos da escala ordinal é nas chamadas medidas de atitude onde se pede a um respondente uma avaliação numérica de sua opinião a respeito de um certo tema, como por exemplo, em relação à legalização do aborto.

É comum identificarmos estas categorias por um código numérico, e desde que preservemos sua ordenação, não há qualquer restrição adicional quanto aos valores a serem utilizados. A prática, no entanto, recomenda que utilizemos valores inteiros e consecutivos a partir da unidade, não importando se esta codificação é feita na ordem crescente ou decrescente.

Assim, no caso do nosso exemplo, poderíamos utilizar as seguintes codificações para classificar as empresas quanto ao tamanho:

(1) Pequena (2) Média (3) Grande , ou

(1) Grande (2) Média (3) Pequena .

Embora o nível ordinal contenha mais informação que o nominal, a magnitude da diferença entre seus valores não reflete uma medida de distância. Desta forma, as medidas estatísticas que fazem uso desta propriedade não são válidas para dados em escala ordinal.

Assim sendo, em princípio ñao se deve calcular a média e o desvio padrão para dados ordinais. Mas esta questão representa um tema até certo ponto polêmico, uma vez que alguns autores admitem o seu uso em determinadas situações, como por exemplo ao se trabalhar com medidas de atitude. Ao fazê-lo, no entanto, recomendamos que o pesquisador mencione este fato em seu estudo, assegurando assim um maior respaldo técnico às suas análises.

Em relação aos dados nominais, uma propriedade adicional que os dados ordinais apresentam é a transitividade. Assim, tendo três valores A, B e C, tais que A > B e B > C , então poderemos afirmar que A > C .

Do ponto de vista estatístico, os dados em escala ordinal admitem as operações que envolvem a contagem de frequências e, naturalmente, a ordenação dos valores. Exemplo de procedimentos estatísticos aplicáveis seriam o cáculo da mediana e as medidas de correlação ordinal.

Nível Intervalar:

Corresponde ao nível de mensuração em que um número já expressa uma grandeza realmente quantitativa. O nível ou escala intervalar é aquele em que, por razões de natureza prática, o valor zero é atribuído para indicar apenas um ponto de referência, não tendo a conotação de ausência de uma propriedade.

O melhor exemplo de escala intervalar é a das temperaturas, em graus centígrados, com que estamos acostumados. Neste caso, não se pode dizer que uma temperatura de 30 graus represente duas vezes mais calor do que uma de 15 graus. No entanto, tem-se que a diferença entre 10 graus e 30 graus é o dobro da diferença entre 80 e 90 graus.

Na escala intervalar as distâncias são grandezas comparáveis. Por isso, a escala intervalar admite praticamente todas as operações aritméticas usualmente utilizadas nos cáculos estatísticos, como por exemplo para a determinação da média e do desvio padrão.

A única operação aritmética não definida para dados intervalares é a divisão ou razão entre seus valores. Esta restrição, no entanto, impõe poucas limitacões em termos de cáculos estatísticos.

Nível de Razão:

Ao contrário da escala intervalar em que o zero é apenas um ponto de referência, na escala de razão o valor zero indica realmente uma ausência da característica medida. Mas, como já foi dito, esta diferenciação tem pouca relevância do ponto de vista estatístico.

A escala de razão expressa resultados de contagens, sendo que, agora, a razão entre dois valores tem significado físico. Exemplos de dados nesta escala são muito comuns: o peso e altura das pessoas, a taxa de câmbio, o índice de inflação, o número de filhos de um casal,...

Sendo esta a mais completa das escalas, todos os procedimentos estatísticos podem ser livremente aplicados aos dados deste tipo. Cabe mencionar, no entanto, que o mesmo praticamente vale para os dados intervalares.

O Caso especial de dados dicotômicos:

Variáveis dicotômicas são aquelas que podem assumir somente dois valores distintos, como na classificação de pessoas quanto ao sexo ou o hábito de fumar. Dados como esses são geralmente do tipo nominal. No entanto, o fato de serem somente duas as classes, permite que os dados sejam arbitrariamente ordenados, podendo ser tratados como se fossem ordinais, ou mesmo intervalares.

Com isso, aumentamos o repertório de análises estatísticas possíveis com este tipo de dado, habilitando-nos a calcular medidas como sua mediana, média e desvio padrão.