

Cap. 20: A ANÁLISE DE DADOS

GOODE, W. e P.K. HATT (1968)
MÉTODOS EM PESQUISA SOCIAL - p.434-455

Até agora demos destaque à necessidade de um planejamento cuidadoso da coleta de dados para que as questões significativas possam ser respondidas, e aos problemas da coleta de dados para que as respostas sejam fidedignas e válidas. É óbvio, porém, que só a realização desses dois passos não trará uma resposta à hipótese original. Ao contrário, proverá meios de respondê-la. Este capítulo visa esclarecer alguns modos pelos quais esses dados podem ser ordenados para dar as respostas desejadas.

Em condições ideais de precisão e simplicidade, isto apresenta somente alguns problemas que a proposição da hipótese e a elaboração do plano experimental darão automaticamente à análise dos dados. Considere, por exemplo, uma hipótese a seguinte: Se luzes são selecionadas como estímulos visuais de maneira que cada uma seja sublimar (não pode ser percebida) para cada olho tomado separadamente, então se são aplicadas simultaneamente (mas ainda separadas) são percebidas pelo sujeito. Esta hipótese está de tal maneira proposta, que poderá haver ou não percepção. A análise dos dados não constituirá problema. Se, por outro lado, as respostas mostrarem não uma presença ou ausência absoluta mas proporções ou graus de presença, o problema da análise tornar-se-ia mais complexo. Se, além disso, as proporções daqueles que percebem a luz parece ser diferente entre tipos diversos de pessoas, por exemplo, homem e mulher, pessoas mais velhas e mais jovens, o problema se torna logo ainda mais complexo.

Em outras palavras, os problemas criados pela análise dos da-

Tamanho da Família	Porcentagem de Famílias Residindo em:		
	Casas pequenas	Casas médias	Casas grandes
Famílias grandes	25,0	55,0	20,0
Famílias médias.....	30,0	60,0	10,0
Famílias pequenas.....	30,0	55,0	15,0

dos estão diretamente relacionados com a complexidade das hipóteses ou hipótese. Suponha, por exemplo, um estudo que é realizado para verificar a hipótese de que o tamanho da família está diretamente relacionado com o tamanho do lar onde reside a família, os resultados aparecem na tabela acima. Estes resultados provocariam rejeição da hipótese tal como foi apresentada e levariam a uma reformulação em termos mais específicos e por isso mais complexos. Assim, poderia ser reformulada da seguinte maneira: O tamanho da casa está positivamente relacionado com o tamanho da família, quando as oportunidades de escolha são iguais. Isto naturalmente exige uma definição clara dos fatores que influenciam a liberdade de escolha. Logicamente, a capacidade financeira de comprar ou alugar nos vários níveis é um desses fatores, já que o tamanho da casa se correlaciona com seu preço. Uma nova tabulação, portanto, como a seguinte pode aparecer. Este tipo de resultado tenderia a concordar genericamente com a hipótese reformulada. Se tabulações como estas forem feitas para famílias "médias" e "pequenas", e estas coincidirem com os resultados abaixo, então a hipótese seria ainda mais amplamente confirmada.

Entretanto, a capacidade financeira não é o único fator que interfere com a escolha de uma casa. Em algumas áreas ser membro

de uma raça "não-caucasiana" inibirá a escolha por meio de contratos restritivos. Portanto, será talvez necessário fazer duas divisões

Nível Econômico da Família	Percentagem de Famílias Grandes Residindo em:		
	Casas pequenas	Casas médias	Casas grandes
Alto.....	5,0	35,0	60,0
Médio.....	35,0	65,0	10,0
Baixo.....	40,0	55,0	5,0

raciais (brancos e não-brancos) dentro de três níveis econômicos (alto médio e baixo) para três tamanhos de famílias (grande, médio e pequeno). Isto daria 2x3x3, ou 18 tabelas. Como se verá mais adiante, nem sempre é necessário realizar a tarefa de modo tão incômo-do, embora se deva considerar o problema lógico.

Se todos os fatores relevantes são conhecidos antes, não existi-rão problemas sérios na análise, pois como foi salientado em capítu-los anteriores, o plano experimental teria colecionado os dados somente nos casos mais importantes. Existem, porém, dois aspectos da pesquisa científica que tornam relativamente rara a existência de uma situação tão ideal. O primeiro dêsses é o aparecimento de uma regularidade empírica anômala ou a ausência de uma regularidade esperada. Êsses casos exigem uma análise não prevista no plano ori-ginal. Semelhante a isto é o uso de dados obtidos para um propósi-to, num problema, bem diferente. Ambos representam o que se deno-mina análise secundária. De um ponto de vista "puramente" experi-mental, essas análises dão respostas que são "plausíveis" mas não podem ser propostas nos termos de "probabilidade", comuns à ciên-cia.

Por exemplo, no caso do estudo de residência mencionado acima, se raça e status econômico foram considerados no plano original, a amostra empregada deveria ser de tal maneira construída para conter uma seleção adequada e representativa de todos os vários subtipos. Isto é, uma amostra satisfatória de cada tamanho de família dentro de cada grupo racial e dentro de cada nível econômico deveria ser obtida. Se isto não foi feito porque essa análise resultou de uma decisão posterior, torna-se difícil julgar do valor da amostra antiga para os propósitos novos.

Porém, a análise secundária não é somente comum mas necessária, e é certamente de grande valor, mesmo se não produzir mais do que afirmações plausíveis para servir como hipóteses para verificação posterior em termos mais rigorosos(1).

Num sentido, portanto, os problemas de análise dos dados envolvem todas as questões propostas no capítulo sobre plano de estudo, pois a análise secundária envolve o planejamento e o replanejamento de substitutos para o experimento controlado. Isto indica que existem dois tipos de questões sobre análise que devem ser apresentadas. Uma delas se relaciona com as técnicas de representar os dados, e a outra com métodos de ordenar logicamente para que as questões possam ser propostas e respondidas.

O primeiro desses problemas gerais não pode ser tratado exaustivamente num livro como este; em todo o caso, é sempre parte do conteúdo comum de um curso de estatística. Serão feitos alguns comentários básicos sobre os métodos de representar os dados, focalizando mais propriamente o processo lógico envolvido na análise secundária.

(1) Uma discussão pormenorizada deste ponto pode ser encontrada em Robert K. Merton e Paul F. Lazarsfeld (orgs.). Continuities

REPRESENTAÇÃO ESTATÍSTICA

A distribuição por frequência. A forma mais simples de representar os dados da pesquisa é a distribuição por frequência ou tabulação. Isto se resume na apresentação numa coluna de qualidades diferentes de um atributo, ou valores diferentes de uma variável, junto com as entradas em outra coluna mostrando a frequência da ocorrência de cada uma das classes. Os únicos problemas relacionados com a preparação de uma distribuição por frequência útil, ou tabela simples como pode ser chamada, é usar o senso comum em relação a três coisas.

Primeira, as unidades colocadas na coluna à esquerda descrevendo as qualidades ou valores devem ser mutuamente exclusivas, bem como incluir a totalidade de observações que serão feitas. Valores sobrepostos ou atributos só podem gerar confusão. Por exemplo, num estudo de Porto Rico um dos autores verificou que além das categorias católico, protestante, e "sem religião", alguns professavam o espiritismo. Isto não causaria problema exceto pelo fato de que parece possível combinar espiritismo quer com protestantismo ou com catolicismo. Portanto, ao invés de simplesmente somar uma categoria, espíritas, foi necessário acrescentar também católicos e espíritas.

Segunda, a tabulação, para ser de máxima utilidade, deve ter lógica interna e ordem. Parece bem óbvio que, se tabulasse a variável altura do homem ou tamanho de cidades, se deveria tabular em ordem a partir do mais alto ou maior para o mais baixo ou menor,

(1)

in Social Research, Free Press, Glencoe, III., 1950, passim, esp. pp. 133-143 e pp. 197-211.

ou vice-versa. Entretanto, quando se tabulam qualidades, quando a ordem não pode ser tão óbvia, a necessidade de um tratamento lógico é igualmente grande. Por exemplo, um survey perguntou a mulheres quais as razões para comprar um determinado creme facial(2). Uma tabulação dos resultados é encontrada na tabela seguinte:

Razões para comprar creme facial (§)

Razões	Percentagem de informantes
Recomendação.....	28,0
Beneficia a pele	21,0
Ouviu o reclame no rádio	18,0
Viu sobre o balcão.....	15,0
Preço razoável	10,0
Bom aroma	8,0
Condições especiais da pele	7,0
Total	107,0

(§) Estritamente falando, isto não é uma frequência mas uma distribuição de percentagem. Este problema será considerado adiante.

Em primeiro lugar, como o total alcança 107 por cento, é evidente que esta tabela viola a primeira exigência, isto é, de que as categorias devem ser mutuamente exclusivas. Ignorando este fato, porém, nota-se que não há uma ordem sensível de tipos de razões apresentadas. Como é possível melhorar a utilidade da tabulação para fins de análise? Como não existe contínuo como o de altura ou peso, o procedimento é procurar agrupamentos lógicos de respostas. É possível descobrir três dessas grandes categorias e o uso da tabela aumenta muito se modificada (ver pág. seguinte).

(2) Adaptado de Hans Zeisel, Say It with Figures, Harper, Nova York, 1947, 2a. ed., pp. 7-8.

Nem tôdas as tabulações têm uma estrutura lógica, mas na maioria dos casos é possível encontrar, mesmo na tabulação simples, uma ordem lógica para usar como um princípio de análise.

Razões	Percentagem de informantes
Relativas ao informante:	28,0
Beneficia a pele 21,0	
Condições especiais da pele..... 7,0	
Relativas ao produto:	18,0
Preço razoavel..... 10,0	
Bom aroma..... 8,0	
Relativas ao meio através do qual ouviu sobre o produto:	61,0
Recomendação..... 28,0	
Ouviu o reclame no rádio 18,0	
Viu sobre o balcão 15,0	

Terceira, quando a coluna à esquerda de uma tabulação é uma variável quantitativa como o tamanho de cidade, ou renda mensal, os intervalos de classe devem ser escolhidos cuidadosa e razoavelmente Schmid sugere três critérios para isso(3):

"1. Ordinariamente não se deve ter menos de 8 ou 10 e não mais de 18 ou 20 intervalos de classe, dependendo da natureza dos dados e do número de casos estudados. Para obter-se uma compreensão clara dos dados originais, os itens individuais são frequentemente arranjados numa ordem de magnitude ascendente ou ascendente. Essa classificação é conhecida como rol. Depois de notar os valores mais alto e mais baixo, bem como os aspectos característicos dos dados, o número de intervalos pode ser mais facilmente determi

(3) Calvin F. Schmid, "Basic Statistical Concepts and Tehniques", in Pauline V. Young, Scientific Social Surveys and Research, Prentice-Hall, Nova York, 1949, pp. 291-292.

nado.

"2. Todo esforço deve ser feito no sentido de obter intervalos de tamanho uniforme. Os intervalos não devem ser tão pequenos para diminuir as vantagens do resumo ou tão grande para esconder as características mais importantes da distribuição. Além disso, se os intervalos de classe são muito pequenos, podem ocorrer intervalos vazios ou em branco. Se se fizer comparações entre lados semelhantes, é aconselhável selecionar intervalos de classe para todas as distribuições. Sempre que possível os intervalos de classe devem representar divisões numéricas comuns ou convencionais como 5 ou 10, e não divisões ímpares como 3 ou 7.

"3. Depois de determinar os intervalos de classe, é importante que eles sejam claramente designados na tabela de frequência. Cada intervalo deve ter limites inferior e superior definidos, e deve ser expresso de maneira a eliminar qualquer possibilidade de má interpretação ou confusão".

Êstes parecem problemas simples, e na realidade o são. Na prática, porém, são um tanto difíceis, e muitas análises de dados foram difíceis de serem realizadas ou compreendidas porque êsses princípios foram ignorados.

Resumindo a distribuição por frequência. Dois são os meios gerais de apresentar de maneira simples uma distribuição por frequência. Êstes são usados isolados ou em conjunto, dependendo do problema à mão. Um dêles é calcular o valor que representa a tendência central das distribuições. Essas medidas são denominadas médias e incluem entre outras a média comum, tecnicamente conhecida como média; a mediana, um valor tal que divide a tabela de frequência em duas partes iguais; e a moda, ou o valor representado pela maior frequência.

O outro tipo geral de resumo de uma distribuição por frequência inclui medidas de dispersão como o desvio-padrão e o coeficiente de variação. Essas medidas são usadas para comparar a amplitude relativa da variação em duas ou mais distribuições por frequência. Suas características, aplicações, e métodos de cálculo podem ser encontrados em qualquer livro de texto elementar de estatística e não serão considerados aqui. São mencionados apenas por serem muito usados como meios de resumir distribuição por frequência.

Como comparar frequências. Um método comum e simples de comparar frequência é por meio da razão (razão por quociente). Uma razão é um mero indicador ou um quociente que relaciona o tamanho de um número a outro. Sua principal utilidade é a de atuar como uma medida relativa e assim permitir a comparação de números também desiguais. Por exemplo, se desejamos conhecer a relação de proporção de mortes femininas e masculinas de várias idades, uma série de razões como mostra a tabela seguinte será útil.

Acrescentam ao conhecimento, como se vê nas segunda e terceira colunas, que a proporção de morte de homens é geralmente maior do que a de mulheres, ao mostrar que esta vantagem feminina é constante através da vida, com exceção de ser um pouco maior do que o comum nas idades de 5 a 9, e um pouco menor nas idades de 75 para mais. Os algarismos na quarta coluna são quocientes; poderiam ser expressos como 172 a 136, 24 a 17, 15 a 12, etc., mas os quocientes são mais úteis, pois reduzem os algarismos à direita a 1 em cada caso, possibilitando comparar mais facilmente.

Um método correlato de comparar valores é a proporção. Esta medida é uma fração cujo numerador é uma das duas frequências observadas e o denominador a soma das frequências observadas. Assim, as tabelas anteriores poderiam ser expressas como proporções de todas as mortes nas várias idades que foram mortes de ho-

Mortes por centena da população branca nativa por idade e sexo, nas cidades de 100.000 habitantes ou mais, Ohio, 1930(\$).

Idade	Homem	Mulher	Razao de homem para mulher
0 - 4	17,2	13,6	1,26
5 - 9	2,4	1,7	1,41
10 - 14	1,5	1,2	1,25
15 - 19	2,4	1,9	1,26
45 - 54	12,3	10,0	1,23
55 - 64	25,8	19,9	1,30
65 - 74	59,1	45,9	1,29
75 e mais	135,6	114,5	1,18

mens. Assim, os quocientes 1,26; 1,41; 1,25, etc., expressos como proporções, seriam 0,558; 0,585; 0,555, etc. Quando as proporções são expressas em múltiplos de 100, elas são percentagens. Assim, no exemplo acima, 55,8 por cento de todas as mortes nas nas grandes cidades de Ohio em 1930, entre as idades de 0 a 4, são mortes de indivíduos do sexo masculino, etc. A escolha entre razão, proporção, ou percentagem para a análise de dados é uma questão de pura preferência e depende da maneira como o pesquisador deseja comunicar seus resultados.

O uso de percentagens (bem como de razões e proporções) tem por finalidade simplificar o problema de comparação. É importante, assim, ver exatamente o que o seu uso significa para não usá-las impropriamente. Primeiro, podem servir para dar forma numérica às características qualitativas.

Assim, é possível comparar duas classes escolares na base do sexo, dizendo que uma classe é 60 por cento masculina e a outra só 40 por cento. Esta é uma comparação perfeitamente boa, mas de

ve-se tomar cuidado para não traduzir esta afirmação dizendo que uma é mais masculina do que a outra e assim supondo que a masculinidade é uma característica quantitativa. Esses erros de lógica são comuns e facilmente cometidos, mas as consequências podem ser de sastrosas.

Segundo, as percentagens reduzem duas distribuições por frequência a uma base comum, como foi acima ilustrado, simplificando muito a comparação. Deve-se notar, porém, que isto oculta comparações absolutas, e às vezes são estas e não as comparações relativas as importantes. Zeisel dá um exemplo desse problema, como se vê na tabela seguinte.

Crescimento comparativo das cidades M e N,
1941 - 1946

Cidade	1941	1946	Aumento	Percentagem de aumento
M	1.000.000	1.200.000	200.000	20,0
N	500.000	650.000	150.000	30,0

Os algarismos relativos e absolutos produzem, aparentemente, resultados diferentes. O problema, então, é o de saber qual o "certo". A cidade M cresceu mais ou menos rapidamente do que a cidade N? É claro que este é um problema lógico. Assim, pode-se raciocinar que, se o crescimento fosse "normal" deveríamos esperar que a cidade maior atraia mais pessoas do que a menor, e assim, as percentagens são algarismos "significativos". Se, por outro lado, o o crescimento das cidades depende não do tamanho anterior mas de outros fatores, como novas indústrias, então podemos dizer que a cidade maior cresceu mais rapidamente em razão de 200.000 para 150.000, ou uma vez e um terço mais rápido.

Em outras palavras, embora as percentagens auxiliem muito a comunicação por simplificação, são também susceptíveis de conduzi-rem a erro por obscurecerem fatos significativos. Isto ressalta a importância de nunca apresentar uma tabela de percentagem sem também mostrar os números brutos que elas representam. Isto não somente previne esse tipo de falsa interpretação descrita acima, mas garante ao leitor que existe uma base adequada para o cálculo da percentagem. Por exemplo, a afirmação de que três de um total de quatro pessoas entrevistadas são a favor do Partido Democrático torna-se muito confusa sob a forma "três em cada quatro" ou "75 por cento" dos entrevistados a favor da plataforma democrática.

Bases para calcular percentagens. A transformação de uma distribuição por frequência simples em percentagem tem, geralmente, por finalidade, esclarecer as comparações entre vários intervalos de classe. Neste caso, muito raramente o conhecimento da base de percentagem constitui problema. Um caso, porém, no qual a dificuldade pode surgir, é quando a tabulação inclui categorias residuais, ou mais de uma dimensão. Um exemplo vai mostrar isto mais claramente. Se se solicita a opinião de uma amostra sobre emissão de ações, e os resultados são apresentados na tabela seguinte, haverá um problema de interpretação. Assim, se a base da percentagem foi considerada a amostra total, aqueles a favor seriam 39 por cento; se a base foi todos os que responderam, o resultado seria 44 por cento; e ainda se fosse considerada como aqueles que responderão e que já se decidiram, seria 56 por cento. É muito importante, como se ve, apresentar completa e cuidadosamente a base da percentagem para evitar confusão.

O problema mais comum surge quando se usa uma tabulação de dupla entrada pois, nesse caso, mesmo se as categorias são completas, é ainda necessário fazer uma escolha para saber que percenta

gens devem ser calculadas na horizontal ou na vertical. Zeisel apresenta uma regra geral, que pode ser frequentemente obedecida, afirmando que "as percentagens devem ser calculadas na direção do fator causal".

Repostas à questão: Você está de acordo com a emissão de ações escolares que serão votadas em novembro?

Respostas	Número
Sim.....	97
Não	78
Indecisos	44
Recusam responder	31
Total	250

Isto não significa que um dos fatores deva ser a causa real do outro, mas apenas que na mente do analista um deles é pensado como influenciando sobre o outro. A ilustração dada por Zeisel esclarece melhor. Primeiro, as percentagens podem ser calculadas em relação aos totais vertical ou horizontal.

Mortes devidas ao câncer nos E.U.A. por raça

RAÇA	CAUSA DE MORTE		
	Câncer	Tôdas as outras	Total
Branca.....	139.627	1.055.804	1.195.431
Negra	9.182	169.391	178.573
Total	148.809	1.225.195	1.374.004

Assim, os números apresentados na primeira destas tabelas podem ser expressos em percentagens calculadas verticalmente, para parecer como se mostram na segunda tabela. Desta forma a apresen

tação mostra a composição racial dos dois grupos, aqueles que morrem de câncer e aqueles de outras causas. Esta é uma maneira de sajeitada de apresentar a análise, pois provavelmente não supomos que a causa de morte seja materialmente afetada pelo número relativo de brancos e negros. Ao contrário, é provável que a hipótese fosse que por várias razões a morte acompanhada por um diagnóstico de câncer é mais comum num grupo do que no outro. Portanto, as percentagens deveriam ser calculadas horizontalmente, como se mostra na terceira tabela. Esse arranjo mostra diretamente a comparação desejada entre brancos e negros, e não exige outra análise como foi o caso da outra maneira de calcular percentagem.

RAÇA	Câncer	Todas as outras	Total
Branca.....	93,8	86,2	87,0
Negra	6,2	13,8	13,0
Total	100,0	100,0	100,0

RAÇA	Câncer	Todas as outras	Total
Branca	11,7	88,3	100,0
Negra	5,1	94,9	100,0
Total	10,7	89,3	100,0

Frequentemente ocorre não aplicar-se o "princípio causal" tão claramente como no caso acima. Por exemplo, suponha que se faça uma tabela de dupla entrada com o número de anos completos de escola e "valor econômico" para uma série de adultos; a maior instrução seria considerada a "causa" do nível econômico mais elevado, ou vice-versa? Naturalmente, poderia ser visto de ambos os ân

gulos. O princípio geral, então, é o de calcular a direção do fator que o analista deseja ressaltar.

O conceito que o analista possui de causalidade ou desejo de dar ênfase não é, porém, sempre o único determinante da direção na qual as percentagens deveriam ser calculadas, como se mostra num outro princípio proposto por Zeisel. Este princípio é o de que a percentagem se deve dirigir somente na direção na qual a amostra é representada. A ilustração dessa regra trata da coleção de preferências políticas numa amostra representativa de 8.000 pessoas em cada um dos dois estudos, cujos resultados aparecem na tabela seguinte. Se esses resultados são analisados do ponto de vista de que as diferenças entre os Estados "causam" as diferentes lealdades ao partido, então se calculariam horizontalmente as percentagens da tabela abaixo. Esses resultados são bem "sensíveis" ao mostrarem que o partido YY tem uma posição favorável nos dois Estados, mas especialmente no Estado A.

Estado	Partido a favor		Total
	XX	YY	
A	2.500	5.500	8.000
B	3.500	4.500	8.000

Estado	Percentagem favorável ao Partido		Total
	XX	YY	
A	31,0	69,0	100,0
B	44,0	56,0	100,0

Se a questão original é invertida, porém, e se se supusesse por

hipótese que as diferenças nos partidos e não nos diferentes Estados eram a causa da distribuição, as percentagens seriam calculadas verticalmente, como mostra a tabela seguinte. Se esta tabela apresentasse uma análise precisa, pareceria dizer que 42 por cento dos votos do partido XX viriam do Estado A. Se for lembrado, porém, que as amostras representam cada Estado e não os dois partidos, é claro que essas percentagens não têm sentido. Como não conhecemos a contribuição relativa dos dois Estados ao total de votos, essa percentagem não pode ser significativa. Portanto, quando as tabulações de dupla entrada são representativas apenas numa direção, as percentagens só podem ser calculadas nessa mesma direção.

Estado	Percentagem a favor do Partido	
	XX	YY
A.....	42,0	55,0
B.....	58,0	45,0
Total.....	100,0	100,0

Ordenação Lógica dos Dados

A tabulação de dupla entrada ou de mais atributos e variáveis é apenas um método formal e econômico de arranjar os dados, de modo a permitir a aplicação de métodos lógicos de verificação. Assim, os métodos de concordância, diferenças, ou variação concomitante (correlação) podem ser usados na obtenção das conclusões de uma tabulação de dupla entrada. Isto deve também salientar o ponto já mencionado neste capítulo, de que o que se pode obter através de tabulação múltipla depende inteiramente do plano lógico do estudo e do discernimento do investigador. Não existem "truques" na tabulação múltipla que possa garantir que uma análise dará os resulta-

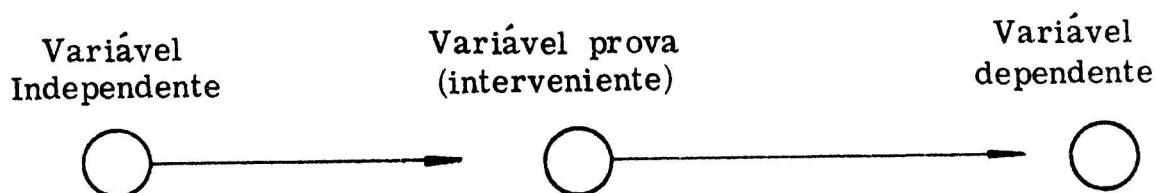
dos mais significativos e significantes possíveis.

É possível, porém, discutir alguns modos de raciocínio que podem conduzir a tabulação múltipla. Antes de mais nada, o uso de tabulação de dupla entrada se aproxima, com efeito, do experimento controlado. Isto significa que o analista está realmente pensando em termos de causa e efeito. Isto é, ele tem em mente uma ou mais variáveis, variação que pode ser usada para explicar a variação de outra variável. Estas dimensões "causais" são denominadas variáveis independentes, e os valores a serem explicados, variáveis dependentes. Como o analista pensa realmente em termos causais é necessário elaborar os resultados do survey com a tabulação múltipla. Se tudo o que o analista deseja é uma afirmação empírica de variação concomitante, a simples tabulação de dupla entrada será suficiente. A necessidade, porém, de conhecer o "significado" de uma relação leva a fazer muitas outras perguntas que só podem ser respondidas introduzindo-se outras variáveis na análise.

Kendall e Lazarsfeld apresentaram uma classificação de três tipos de elaboração que eles denominaram interpretação, explicação e especificação.

Interpretação. Este é um processo de estratificar uma tabulação de dupla entrada anterior com outra variável (chamada variável-prova) quando a variável-prova ocorreu 1) depois da variável independente, mas 2) antes da dependente. Uma outra exigência é a de correlacionar o fator-prova com ambas as variáveis dependentes e independente. Esse procedimento permite uma nova interpretação de uma relação original em termos de um terceiro fator, neste caso denominado variável interveniente.

Isto pode ser representado com o seguinte diagrama:



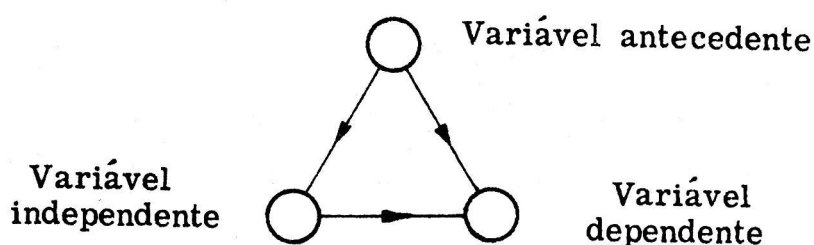
Por exemplo, quando os estudantes masculinos de uma universidade particular foram tabulados de acordo com a escola secundária pública ou particular frequentada e pelo seu boletim acadêmico universitário, aqueles que frequentaram escolas públicas mostraram um boletim superior. Diante disso, esse resultado sugere uma explicação da variável dependente (notas) em termos de excelência da preparação na variável independente (escolas secundárias públicas versus particulares). Qualquer pessoa familiarizada com a política de admissão das universidades privadas pode suspeitar, porém, que fatos ocorreram entre treino secundário e aceitação na universidade, que podem dar aos resultados um significado inteiramente diferente.

Assim, uma variável interveniente, como excelência de trabalho acadêmico na escola secundária poderia ser logicamente introduzida. Se circunstâncias existiram no procedimento de inscrição, significando que os formados por escola pública enfrentaram padrões acadêmicos mais elevados do que os formados por escola particular, a correlação original entre o tipo de escola secundária e graus subsequentes obtidos desaparecia. Neste caso, o fator-prova, boletim de notas do secundário, serviria para interpretar os resultados originais, pois quando homens do mesmo nível de realização escolar são comparados, desaparece a relação entre o tipo de escola e êxito no secundário.

Explicação. Este é essencialmente o mesmo tipo de raciocínio da interpretação no qual se procura reduzir uma correlação origi-

nalmente observada através do uso de um fator-prova. Neste caso, porém, o fator-prova é uma variável antecedente e não uma interveniente. Em outras palavras, uma relação observada é explicada em termos de um terceiro fator que ocorreu antes da variável independente e que está relacionado com ambas.

Isto pode ser representado com o seguinte diagrama:



Por exemplo, um estudo da taxa de nascimento em Pôrto Rico apresentou uma tabulação de dupla entrada que mostrou existir uma correlação negativa entre status sócio-econômico e taxa de nascimento. Ambas as variáveis nesse caso eram medidas da experiência de vida do adulto, e como o raciocínio "causal" foi envolvido se perguntou se existia ou não alguma variável caracterizando uma parte inicial da história de vida que pudesse "explicar" essa correlação.

Assim a instrução foi usada como uma variável-prova antecedente, verificando-se que a correlação entre status sócio-econômico e taxas de nascimento quase desapareceu quando tabulada separadamente em relação a níveis de instrução comparáveis. Neste caso, então, o uso de um fator-prova "eliminou" uma original e a substituiu por outra. Descreveríamos a relação entre status sócio-econômico e taxa de nascimento em Pôrto Rico como uma correlação espúria. Isto levanta a questão. "Como podemos ter certeza de que a relação entre educação e taxa de nascimento é uma causa real e não apenas outra correlação ilegítima?" A única resposta para isto é dizer que se a aplicação de todos os fatores-prova relevantes não reduz mate

rialmente essa correlação, pode-se supor a causalidade.

Especificação. Enquanto na explicação e interpretação a questão é a de saber se a variável-prova causará uma clara redução ou desaparecimento de uma relação observada, o objetivo da especificação é bem diferente. Neste caso a sequência "causal" é modificada especificando condições várias sob as quais a correlação original existirá com maior ou menor intensidade. O exemplo dado por Kendall e Lazarsfeld usa resultados do The American Soldier que aparecem na tabela seguinte.

Posição	Diploma da High School ou mais	Menos do que High School
Non-com.....	61,0	43,0
Pvt., Pfc.....	39,0	57,0
Total de casos.....	3.222	3.152

Uma medida bruta da relação entre educação e posição é a diferença, entre 61 por cento e 43 por cento. Este valor, chamado f , é 0,18. Neste ponto, se perguntou se um valor de f de 0,18 seria provavelmente encontrado em todas as circunstâncias; pareceu que a resposta é "não". Isto é, oportunidades para promoção podem não ter sido iguais durante todos os períodos da Segunda Grande Guerra. Por isso foi feita uma tabulação de dupla entrada com a duração de serviço, como aparece na tabela abaixo.

Posição	Serviu menos de dois anos		Serviu dois anos ou mais	
	Diplomada High School ou mais	Menos do que High School	Diploma da High School	Menos do que High School
Non-Com.	23,0	17,0	74,0	53,0
Pvt., Pfc.	77,0	83,0	26,0	47,0
Número to- tal de casos	842	823	2.380	2.329
	f = 0,06		f = 0,21	

Como os f são bem diferentes para as duas categorias de tempo de serviço, o analista pode agora afirmar que a relação entre instrução e posição é maior para aqueles que entraram cedo para o exército do que para os que entraram tarde.

Especificação, portanto, é o processo de descrever as condições nas quais uma determinada inter-relação pode ou não existir, ou pode existir num grau maior ou menor. Como a explicação e a interpretação, portanto, torna a análise do "causa" mais aceitável e significativa.

Aspectos gerais da elaboração. Dois problemas mais importantes parecem existir sobre este aspecto da análise. A aplicação dos fatores-prova nestes três parágrafos precedentes sem dúvida parece bastante simples. Os dois problemas práticos, porém, são 1) elaborar ou não, e 2) que variáveis-prova aplicar.

Quanto ao primeiro problema, pode-se indicar que consequências não esperadas podem levar o analista a elaborar. Se os resultados concordam com a expectativa, existe pouco estímulo para análise ulterior. Esta é uma situação também perigosa pois pode terminar com a aceitação de relações ilegítimas sem interpretação.

Como Kendall e Lazarsfeld afirmam:

" Nossa primeira preocupação é saber se a relação é ou não legítima; conseqüentemente nossos esforços iniciais de elaboração são geralmente do tipo exploratório. Quando ganhamos alguma garantia de que a relação original não é espúria, tentamos interpretar os resultados ou especificá-los. Perguntamo-nos que variáveis poderiam provocar as relações entre a "causa" e o "efeito", ou que condições poderiam mostrar a relação original mais pronunciada do que originalmente foi verificada. A elaboração de um determinado resultado pode continuar quase indefinidamente. Os únicos limites são a nossa falta de engenhosidade em imaginar os fatores a partir dos quais elaborar os resultados, pela ausência de dados para verificar a importância dos fatores que imaginamos, ou pelas dificuldades de tratar alguns casos à medida que se estende o processo de elaboração"

A elaboração é, portanto, um processo limitado somente pelo analista de um lado: sua capacidade, paciência e objetivos; e, de outro, pela natureza dos seus dados.

Infelizmente, não é possível dar instruções precisas para selecionar os fatores a serem usados como variáveis-prova. Deveria ser lembrado que a seleção de cada variável-prova realmente constitui a formação de uma hipótese que pode ser verificada com os dados à mão. Portanto, aqui se aplica todo material dos capítulos sobre hipótese e plano experimental.

Em outras palavras, a engenhosidade do analista coadjuvada com o conhecimento do campo no qual está trabalhando será a principal fonte de variáveis-prova. Dois métodos de induzir tais hipóteses, porém, podem ser mencionados. Primeiro, seria possível, em alguns estudos, tabular cada fator com os demais fatores como rotina. Na prática isto é geralmente incômodo e o número deve ser reduzido substancialmente tomando-se como base a teoria existente. Em qual-

quer caso, grande número de tabulações de dupla entrada frequentemente sugere a seleção da variável-prova. Segundo, o estudo dos casos discrepantes geralmente também indicará o tipo de interpretação, explicação, ou especificação que irão melhorar a análise.

No sentido mais geral, deve-se dizer que todo o processo de análise não é tanto uma questão de técnicas manipulativas quanto de aplicação rigorosa dos princípios básicos do método científico. O pesquisador que conhece bem os problemas do planejamento da pesquisa terá poucos problemas na análise dos dados.