REVISTA BRASILEIRA DE ESTATÍSTICA

Órgão oficial do IBGE
e Sociedade Brasileira de Estatística

PUBLICAÇÃO TRIMESTRAL

Pedidos de assinatura anual e número avulso ou atrasado para:

Diretoria de Divulgação: Av. Brasil, 15.671 — Lucas — Rio de Janeiro — Brasil
CEP — 21.241
Tel.: 391-7788

Livraria do IBGE: Av. Franklin Roosevelt, 146/loja — Centro — RJ — Brasil
CEP — 20.021
Tel.: 220-9147 / 220-8163
DDD: 011

SUMÁRIO

R. bras. Estat.	Rio de Janeiro	v. 41	n.º 162	p. 169 a 334	abr./jun. 1980

Revista Brasileira de Estatística / Fundação Instituto Brasileiro de Geografia e Estatística. — Rio de Janeiro : IBGE, 1940, jan./mar. (A.1, n.) —

Trimestral.

Órgão oficial do IBGE e Sociedade Brasileira de Estatística.

Variações na denominação do editor : Instituto Brasileiro de Geografia e Estatística, Conselho Nacional de Estatística, Diretoria de Documentação e Divulgação 1936-1967. — Fundação Instituto Brasileiro de Geografia e Estatística, Instituto Brasileiro de Estatística, Diretoria de Documentação e Divulgação, 1967-1969. — Fundação Instituto Brasileiro de Geografia e Estatística, Instituto Brasileiro de Estatística, Departamento de Divulgação Estatística, 1969-1973. — Fundação Instituto Brasileiro de Geografia e Estatística, Departamento de Divulgação Estatística, 1973-1977. — Fundação Instituto Brasileiro de Geografia e Estatística, Diretoria de Divulgação, Centro Editorial, Departamento de Editoração, 1977-

Substitui "Revista de Economia e Estatística" do Serviço de Estatística da Produção, Ministério da Agricultura, 1936, jul(v. 1)-1939, abr.(v. 4). — Mensal.

Apresenta índices anuais e índices acumulados nos v. 25(v. 22-24, 1961-1963); v. 27(v. 25-26, 1964-1965); v. 29(v. 27-28, 1966-1967)

1. Estatística — Periódicos. I. IBGE.

IBGE. Biblioteca Central
RJ-IBGE/78-18

CDD 310.5
CDU 31(05)

Artigos

APROVEITAMENTO E MELHORIA DAS ESTATÍSTICAS VITAIS NO BRASIL

Robert Robichez Cassinelli *
Luís Antônio Pinto de Oliveira

SUMÁRIO

APRESENTAÇÃO

O levantamento das estatísticas do Registro Civil brasileiro passou, a partir de 1974, a ser feito pelo IBGE.

* Chefe da Divisão de Desenvolvimento Metodológico e Projetos Especiais — DIMPE/DESPO/SUEGE.

Contaremos, então, com boas informações sobre casamentos civis e sobre óbitos ocorridos nos principais centros urbanos do País, cruzados por sexo, idade, lugar de residência, mês de ocorrência, etc.

Não se deve, no entanto, ser otimista quanto à qualidade das informações sobre nascimentos na maioria dos casos, e tampouco para óbitos ocorridos fora dos principais centros urbanos. Para esses propomos, a seguir, soluções alternativas, formando um Sistema de Estatísticas Vitais (SEV) apto a produzir dados fidedignos, mesmo a nível nacional.

Desejamos destacar a importância da assessoria técnica que nos foi prestada por Richard Irwin, do U.S. Bureau of the Census. Sua competência e determinação foram decisivas em todas as fases do presente trabalho.

RESUMO

Utilizando-se dados de uma pesquisa piloto realizada no município de Bocaiúva, o problema de obtenção de estatísticas vitais fidedignas é examinado comparando-se três diferentes fontes de dados (registro civil, hospitais e informantes locais).

Os dados sobre nascimentos em áreas rurais foram obtidos a partir de um "arquivo de gestantes" produzindo resultados muito promissores.

A proporção de nascimentos ocorridos em hospital e a proporção de nascimentos registrados, segundo a PNAD 73, são apresentadas, e propõe-se que nas grandes cidades a coleta de dados sobre nascimentos seja feita diretamente dos hospitais.

É também sugerido que a cada cinco anos seja realizada uma pesquisa para avaliar a cobertura dos dados sobre nascimentos e óbitos.

1. GRANDES NÚCLEOS URBANOS: O HOSPITAL COMO FONTE DE DADOS SOBRE NASCIMENTOS

A principal característica deste método consiste em atribuir aos hospitais e aos médicos a responsabilidade primordial da declaração de nascimentos, tendo os pais e parentes responsabilidade apenas secundária. Difere radicalmente, portanto, do enfoque utilizado insistentemente em países subdesenvolvidos, com absoluto fracasso do ponto de vista estatístico.

Naturalmente só se pretende utilizar os hospitais como fonte de dados sobre nascimento nas áreas em que a proporção de nascimentos ocorridos em hospital for grande.

1.1 Proporção de nascimentos ocorridos em hospital

Na PNAD-1973, às mulheres que declararam ter tido filhos nascidos vivos nos 12 meses anteriores à pesquisa foi indagado se a criança nasceu em hospital, se foi registrada e se foi submetida ao batismo ou a alguma outra cerimônia religiosa. Vamos, então, examinar alguns dos resultados obtidos.

BRASIL

TOTAL	2.863.015 (100%)
Registrado	1.691.192 (59%)
Não registrado, não sabe, não declarado	1.171.823 (41%)
Em hospital	1.547.902 (54%)
Fora de hospital, não sabe, não declarado	1.315.113 (46%)

BRASIL URBANO

TOTAL	1.562.304 (100%)
Registrado	1.110.573 (71%)
Não registrado, não sabe, não declarado	451.731 (29%)
Em hospital	1.185.102 (76%)
Fora do hospital, não sabe, não declarado	377.202 (24%)

FONTE — PNAD 1973.

Vemos, então, que no quadro urbano a proporção com declaração "em hospital" já era, em 1973, maior do que "registrado". Observe-se também que esta classificação urbano/rural baseia-se em critérios administrativos, incluindo uma grande quantidade de pequenas cidades e vilas desprovidas de hospitais e que, para os fins do presente estudo, deveriam estar na categoria "cidades pequenas e áreas rurais".

Temos os seguintes resultados a nível regional.

QUADRO URBANO

	Total	*Em Hospital*	*Registrado*
Região I	193.443 (100%)	170.764 (88%)	147.893 (76%)
Região II	365.344 (100%)	331.290 (91%)	333.853 (91%)
Região III	200.376 (100%)	164.405 (82%)	168.968 (84%)
Região IV	177.017 (100%)	130.413 (74%)	142.389 (80%)
Região V	469.931 (100%)	285.480 (61%)	218.851 (47%)
Região VI	22.582 (100%)	20.986 (93%)	18.784 (83%)
Região VII	133 611 (100%)	81.764 (61%)	79.835 (60%)

FONTE — PNAD 1973.

O que nos leva às seguintes conclusões:

1. A proporção "registrado" era maior do que a proporção "em hospital" nas regiões II, III e IV. Entretanto, uma vez que existe sanção para o não registro dentro do prazo e a pesquisa foi feita por órgão do governo, é provável que as proporções de crianças registradas estejam tendenciosas para mais. Numa pesquisa[1] feita no Rio Grande do Sul em 1974 para estimar o sub-registro de nascimentos e óbitos observou-se o seguinte: dos 587 nascimentos pesquisados, 38 foram declarados como tendo sido registrados, entretanto, somente 384 eventos (65%) tiveram seu registro comprovado seja através da certidão de nascimento ou de posterior confirmação no cartório.

2. As proporções "em hospital" para as regiões VI (Distrito Federal) e II (Estado de São Paulo) são bastante altas, apesar de que neste último estão incluídas várias cidades pequenas. Nas outras regiões é necessário excluir as cidades desprovidas de serviços hospitalares.

1.2 O Registro Seletivo

Os dados sobre nascimentos ocorridos em hospital estarão livres da tendenciosidade de seleção que ocorre no registro onde os pais dão preferência a algumas crianças sobre outras. Parece haver, por exemplo, maior interesse em registrar os meninos do que as meninas, segundo nos mostram alguns dados para o Estado de São Paulo, onde vemos que o registro tardio é mais comum nas crianças do sexo feminino e que, ao contrário do que se poderia esperar, a seletividade é maior no município da capital.

Além do sexo, o registro parece ser seletivo em outras categorias também. Uma vez que não tem existência legal, o óbito fetal deixa de ser registrado com mais freqüência do que o nascido vivo, o mesmo se dando com crianças de curta duração de vida.

1.3 Classificação dos Eventos

Nos hospitais a classificação dos eventos (se é nascido vivo ou óbito fetal, por exemplo) é feita por profissionais de saúde e, portanto, mais habilitados a fazê-lo do que o pai da criança. Além disso, trata-se de nascimentos realmente ocorridos e não de registros efetuados, ficando excluídos, entre outros registros espúrios, os registros múltiplos.

[1] Rio Grande do Sul. Secretaria da Saúde. Equipe de Estatística. "Sub-registro de nascimentos e óbitos no Rio Grande do Sul". Porto Alegre, 1976.

TABELA 1

NASCIDOS VIVOS, SEGUNDO O LUGAR DE NASCIMENTO E SEXO

SÃO PAULO — 1971

LUGAR E SEXO	NASCIDOS VIVOS		
	Total	Em anos anteriores	
		Números absolutos	%
ESTADO			
Masculino	253 386	23 571	9,3
Feminino	240 745	28 073	11,7
MUNICÍPIO DE SÃO PAULO			
Masculino	81 708	5 631	6,9
Feminino	77 677	8 047	10,4
INTERIOR			
Masculino	171 678	17 940	10,4
Feminino	163 068	20 026	12,3

FONTE — Secretaria de Planejamento — Estado de São Paulo — Departamento de Estatística — Divisão de Estatística Demográfica.

1.4 Sistemática Operacional

Uma vez determinadas as áreas onde a coleta de nascimentos seria feita através dos hospitais, é imprescindível conhecer a proporção de nascimentos que ocorrem em hospital nessas áreas, ao menos para o total [2]. Avaliações do tipo da PNAD-73 são extremamente úteis, mas ainda insuficientes. Seria necessário um confronto caso a caso entre esta e as informações fornecidas pelos hospitais.

1.5 Perspectivas

O Departamento de Estudos de População (DESPO) do IBGE [3] tem mantido contatos informais com as Secretarias de Saúde de Minas, Paraná e Santa Catarina, com vistas à obtenção de dados sobre nascimentos ocorridos em hospitais que, entretanto, difere um pouco daquele por nós preconizado.

[2] Não só no caso de nascimentos coletados em hospital, mas qualquer sistema de obtenção de dados. Nenhuma serventia têm os dados sobre nascimentos do Registro Civil enquanto não se conhecer ao menos o grau de cobertura dos mesmos.

[3] Antigo Centro Brasileiro de Estudos Demográficos (CBED).

Em nossa opinião, as estatísticas de nascimento no Brasil serão algum dia coletadas nos hospitais, só não podemos precisar quando isto finalmente ocorrerá.

2. A PESQUISA DE CAMPO: A EXPERIÊNCIA PARA OBTENÇÃO DE ESTATÍSTICAS VITAIS EM CIDADES PEQUENAS E ÁREAS RURAIS

A análise da situação das estatísticas vitais em áreas predominantemente rurais deixa claro que tanto para nascimentos como para óbitos os resultados fornecidos pelo Sistema do Registro Civil são bastante precários. As razões das dificuldades encontradas prendem-se a uma multiplicidade de fatores que atuam sobre a propensão da população a fazer a declaração, cabendo unicamente aqui assinalar que, a partir do momento em que passamos a encarar cuidadosamente a situação das estatísticas vitais, impressionou-nos desfavoravelmente o fato de que a declaração do evento, embora legalmente obrigatória, dependia da vontade dos familiares, traduzida em atos pelos deslocamentos dos mesmos ao cartório. Com efeito, se a dependência da iniciativa dos familiares já constitui um problema em áreas de concentração populacional como os centros urbanos, com muito mais razão ela representará um obstáculo para boas estatísticas em áreas de grande dispersão, onde o isolamento é não apenas físico como principalmente econômico e cultural. As conclusões que podem decorrer desta formulação devem ressaltar a necessidade de, ao se proceder a pesquisas experimentais com vistas à obtenção de estatísticas vitais, adotar-se métodos capazes de centralizar o próprio processo de obtenção das informações.

Os passos iniciais da pesquisa foram realizados em condições as mais exploratórias, sendo o objetivo principal a tomada de contato com as situações reais encontradas em áreas do interior do País. Durante as primeiras viagens buscávamos visitar os hospitais e uniddes sanitárias públicas e privadas, os cartórios de Registro Civil e os cemitérios públicos e privados, procurando identificar pontos que servissem ao desenvolvimento da pesquisa. A visita aos cartórios confirmou imediatamente a idéia que se faz sobre as dimensões do sub-registro em áreas predominantemente rurais e foi possível, através de contatos informais com funcionários desses cartórios e também com médicos e profissionais de saúde, aprofundar as hipóteses gerais sobre as causas do problema. Nesta etapa, as conclusões foram de que uma cobertura aceitável das estatísticas vitais através do Registro Civil nestas áreas não é viável em um prazo próximo. Um aspecto positivo foi observado no Registro de Óbitos, que aparentemente para a população da sede de alguns municípios visitados apresenta um grau de cobertura, em princípio, aceitável. A razão parece residir no fato de que os cemitérios na sede são mais

"oficiais", existindo um agente público que é responsável e que mantém o cemitério normalmente fechado, resultando então que qualquer sepultamento está, idealmente, sujeito ao processo de legalização do evento, através do registro, para que seja permitido o enterramento no local. É desnecessário frisar que fora da sede do município tal procedimento inexiste.

No caso dos hospitais e unidades sanitárias verificou-se, de maneira geral, que seus serviços atingiam preferencialmente a área constituída pela sede e povoados próximos, tornando-se progressivamente menos numerosos à medida que os povoados eram mais afastados. Quanto aos cemitérios, a impressão resultante foi de que, à exceção dos cemitérios localizados na sede e em algumas vilas distritais, eles podem ser encontrados com relativa facilidade em cada área ou povoado fisicamente delimitado e são de domínio público (portanto não podem ser designados como "clandestinos"), mas não sofrem qualquer tipo de fiscalização ou controle, estando inteiramente abertos para qualquer sepultamento sem necessidade de nenhuma providência legal anterior. Os cemitérios, em geral, podem ser localizados na maioria dos povoados, havendo em algumas regiões cemitérios exclusivamente para criancinhas, denominados de "cemitérios de anjinhos". Existem ainda pequenos cemitérios, chamados "cruzeiros", localizados em fazendas de propriedades particulares, que muito freqüentemente são utilizados para o sepultamento de crianças, empregados ou agregados e mesmo, em alguns casos, para membros da família do proprietário.

As conclusões preliminares formuladas após as primeiras viagens de observação levou-nos a cogitar que, a partir de uma combinação flexível dos elementos disponíveis (hospital, cartório e cemitérios) e adotando-se o critério de pesquisar formas capazes de dispor de pessoas que centralizassem informações em cada área, tornar-se-ia possível levar à prática uma experiência de campo com vistas à obtenção de estatísticas vitais.

Os passos seguintes, que reconstituem o caminho prático e metodológico adotado, podem sintetizar as particularidades e o alcance científico da pesquisa de campo.

2.1 Descrição das Áreas de Estudo

A seleção das áreas de estudo não esteve ligada a critérios rigorosos do ponto de vista estatístico e sim a sugestões baseadas em boa dose de bom senso, representatividade genérica e informações técnicas de outros órgãos interessados. A condição inicial era de que, em qualquer caso, fossem tomados como área de pesquisa os municípios com maior proporção de população rural. A própria idéia de se tomar o município como unidade de trabalho mereceu algumas considerações, pois, alter-

nativamente, seria possível usar áreas rurais quaisquer, sem necessidade de delimitação rigorosa. Entretanto, para a finalidade de avaliação dos dados obtidos durante o período de pesquisa, fazia-se necessário que esta abrangesse uma área completa em termos de fontes estatísticas disponíveis, o que, no caso, corresponde ao município como unidade mínima.

O segundo critério fundamental de seleção observou um forte componente de bom senso técnico aliado a princípios genéricos de representatividade. Fazia parte dos planos iniciais que à Região Nordeste correspondesse dois municípios, o que acabou não sendo possível. Na Região Sudeste escolheu-se um município do norte de Minas Gerais, localizado na zona mineira da SUDENE, o qual já pode ser considerado como pertencente a uma região de transição entre o Sudeste e o Nordeste. A partir de sugestões da Secretaria de Saúde do Rio Grande do Sul, selecionamos igualmente um município do norte do Estado, em região apontada como de más estatísticas vitais.

Os três municípios possuem as seguintes características gerais:

a) Bocaiúva — situado no norte de Minas Gerais, em região seca com características de transição entre o Sudeste e o Nordeste. Sua população em 1960 era de 31.860 habitantes e em 1970 passou para 35.702, com uma proporção de população rural da ordem de 67,2%. Na década de 60 a sua taxa de crescimento populacional foi baixa, o que sugere estar havendo emigração, provavelmente para a pólo de Montes Claros (60 km de distância) ou para a Grande Belo Horizonte (370 km de distância). Possui 5 distritos, todos com cartórios de Registro Civil e um hospital da Fundação SESP na sede, que em 1970 tinha quase 10.000 pessoas em seus limites urbanos. O município é bastante extenso (5.733 km²) servido em suas ligações por estradas municipais de terra que, às vezes, cobrem grandes distâncias. A cidade é predominantemente comercial e administrativa, enquanto a zona rural combina pequenas propriedades agrícolas (mandioca, milho, feijão, banana, etc.) com propriedades maiores com criação de gado, existindo extensões enormes de terras improdutivas. O nível social e econômico da população rural é bastante baixo, parecendo vigorar um sistema de auto-subsistência complementado aos sábados com pequenas vendas e trocas na feira semanal. Alguns povoados estão condenados a um grande isolamento em face das dificuldades de transporte e o abandono das terras.

b) Jacutinga — situado no Norte do Rio Grande do Sul, é um município pequeno (352 km²) com pouca população (7.061 habitantes em 1970 contra 6.463 habitantes em 1960). Possui 2 distritos, sendo que a sede tinha, em 1970, 668 habitantes em sua área urbana. É portanto um município eminentemente rural, caracterizado pelas pequenas propriedades típicas da colonização estrangeira no sul do País. As culturas agrícolas da região são as da soja e do trigo, as quais proporcionam

regularmente um rendimento monetário aos proprietários. Existem nos povoados formas costumeiras de associação comunitária encontradas nessas chamadas regiões de colônia, e todos os indícios encontrados sugerem uma relativa estabilidade social e econômica. O pequeno crescimento da população na década de 60 pode ser explicado pela saída de filhos de proprietários que vão buscar estudo ou trabalho em outros municípios ou igualmente pela saída de trabalhadores sem terra, visto que parece ser bastante reduzido o número de empregados rurais trabalhando nas pequenas propriedades. As condições de vida em Jacutinga são nitidamente superiores às de Bocaiúva, e sendo as distâncias bem menores, inexistindo povoados ou propriedades efetivamente isoladas, o acesso ao hospital da sede (particular, mas com convênio com INPS e FUNRURAL) é bastante facilitado, o que contribui para o nível geral de saúde.

c) Bacabal — localizado em região central do Maranhão, é um dos maiores municípios do estado em termos de população, possuindo 70.233 habitantes em 1970 contra 79.476 em 1960. É evidente a perda de população, o que parece explicável pela substituição das atividades produtivas na região, a qual libera mão-de-obra, e também aos graves níveis de exploração da mesma. A tradicional cultura do arroz que era feita em terras da União pelos chamados posseiros está sendo rapidamente absorvida por fazendeiros locais e principalmente de outros estados, que estão comprando e cercando as terras para a criação de gado. Tal processo, que já vem se desenvolvendo há uns 10 anos, tem agravado a situação de milhares de famílias camponesas em todo o Maranhão e é muito comum no município de Bacabal. As condições de vida são extremamente baixas, a população rural tem sua principal fonte de sustento no babaçu que fornece uma impressionante lista de produtos (até as casas são cobertas com as folhas da palmeira) e a quebra do coco propicia um pequeno rendimento monetário. O município tem mais de uma centena de povoados espalhados pelos seus 1.609 km² de área e grande parte deles vivem em precárias condições de acesso grande parte do ano, quando a estação das chuvas interrompe os caminhos. Na sede, com 42,16% da população do município, já existe um elevado contingente de população periférica vivendo de atividades ocasionais e normalmente subocupada. O hospital é particular e embora mantenha convênios com os institutos de previdência, não parece estar em condições de contribuir de forma desejável para a melhoria das condições de saúde da população.

2.2 Unidade Geográfica de Trabalho de Campo

Tendo sido convencionado que a pesquisa abrangeria municípios como universo de trabalho, a experiência inicial demonstrou que era necessário desagregar os municípios em unidades menores para que o

trabalho fosse fisicamente realizável. A unidade mínima existente, que é uma característica mais ou menos geral de toda a organização espacial de áreas rurais no Brasil, é o povoado. Sobre este a Fundação IBGE emite a seguinte definição operacional, constante no seu "modelo de levantamento de localidades existentes":

> "Localidade existente no município onde houver um aglomerado permanente de população que, sob o nome de 'povoado', 'arraial', 'lugarejo', 'aglomerado', 'industrial', 'agropastoril', etc., portanto, sem categoria de sede de circunscrição administrativa, tenha se constituído em torno de um número de edificações residenciais com vínculo religioso (em torno de igreja ou capela), comerciais (mercado, feira ou casa comercial), ou industriais (grandes usinas ou fábricas) ou mesmo agropastoris (grandes estabelecimentos agrícolas). Não se incluem as casas residenciais ou comerciais isoladas, exitsentes à margem de estradas na zona rural, bairros e subúrbios de cidades e vilas."

O que caracterizaria, por conseguinte, o povoado seria o aglomerado de população constituído em torno de, freqüentemente, pequenas casas de comércio, capela, grupo escolar e acrescentaríamos, também, um cemitério. Além disso, poderíamos destacar como características definida o fato de existir um nome para a região abrangida para o povoado e de a população do mesmo ter consciência clara de residir em uma área identificada por este nome e saber, em linhas gerais, delimitar a área ocupada por seu povoado em confronto com povoados vizinhos.

A unidade geográfica "povoado", em termos sociais, compreende uma comunidade rural com sentido de auto-identificação e laços culturais, econômicos e até familiares bem estreitos, embora não possa ser entendida como comunidade fechada. O povoado não deve ser visto como refúgio de valores "tradicionais" — mesmo porque as formas de penetração de usos econômicos e valores sociais capitalistas "modernas" fazem se sentir de alguma maneira mesmo em regiões rurais distantes — e sim como meio físico onde a população preserva relações de contato e conhecimento pessoal que, como veremos em seguida, foi o fator determinante para sua inclusão na pesquisa.

2.3 O Sistema de Informantes Estatísticos

Após as visitas iniciais que serviram para conhecimento da organização e da estruturação dos municípios selecionados, observou-se como prioritário o mecanismo de pesquisa a ser experimentado, tomando-se o povoado como unidade física de trabalho. O projeto de pesquisa visava a um acompanhamento periódico, não podendo pois ser confundido como pesquisa domiciliar. Na verdade, tratava-se de sistematizar formas

novas e/ou alternativas de coleta de estatísticas vitais, formas estas que deveriam ser submetidas a uma pesquisa intensiva para avaliação de sua eficácia. Como já assinalamos anteriormente, um dos pontos mais frágeis do método tradicional de coleta de estatísticas vitais (o Registro Civil) reside no fato de a obrigação da declaração ser atribuída a pessoas físicas, geralmente familiares. No sistema empregado na atual pesquisa descartamos imediatamente este procedimento e optamos por experimentar um sistema de informantes estatísticos, a nível de povoado.

A tarefa a ser desempenhada por tais informantes consistiria basicamente em, ao tomar conhecimento de um evento vital em seu povoado, preencher um material de coleta já distribuído e aguardar nossas visitas periódicas para o recolhimento do material. Evidentemente, alguns pressupostos e regras tinham de ser bem definidas preliminarmente. Assim é que o informante deveria comprovadamente conhecer as pessoas residentes em seu povoado, deveria saber ler e escrever e, mais importante, deveria ser capaz de ser aceito pelas pessoas quando fosse fazer indagações sobre características demográficas do evento. Tais requisitos estão normalmente associados a pessoas que, por algum motivo, exerçam certo tipo de liderança comunitária. A professora do grupo escolar é destacadamente a pessoa mais indicada para a tarefa, visto que na grande maioria desses povoados pode ser encontrada uma pequena escola municipal, residindo a professora a maior parte das vezes no próprio povoado. Há casas em que a escola cobre a área de dois povoados vizinhos, o que não altera o esquema geral.

Entretanto, nem sempre as professoras residem no povoado (algumas residem na sede ou nas vilas distritais) e também em outros povoados a escola está fechada ou há falta de professoras. Em tais casos, outros informantes com características aceitáveis devem ser buscados, como comerciantes locais ou moradores capazes de preencher os requisitos. Como veremos adiante, para a coleta de estatísticas de óbitos existe a possibilidade de outros informantes.

O treinamento dos informantes não foi realizado como idealmente se pretendia, ou seja, com uma reunião geral dos mesmos em cada município, o que emprestaria um caráter mais motivador de equipe. As circunstâncias em que a pesquisa foi obrigada a se desenvolver, à base do empreendimento individual de técnicos do DESPO e sem acesso a recursos materiais mais específicos, condicionou-nos atingir o treinamento a explicações sobre a finalidade da pesquisa e à maneira de preencher o material, na própria residência do informante. Cabe assinalar que uma grande limitação da pesquisa é o fato de que os informantes são inteiramente voluntários e o único benefício que auferem é a duvidosa satisfação de estarem colaborando com uma iniciativa de um órgão bastante conhecido como o IBGE. Considerando-se o baixo nível de rendimento monetário que prevalece nestas regiões, em que

mesmo as professoras municipais dificilmente ultrapassam o rendimento mensal de Cr$ 200,00, é natural que os mesmos aspirem a obter algum tipo de remuneração pela atividade exercida.

A substituição de informantes, seja por migração, ausência ou desempenho fraco, é um fato que ocorreu poucas vezes durante a pesquisa, não se revestindo de maiores dificuldades e demandando unicamente uma avaliação criteriosa sobre a pessoa em condições de efetuar a substituição. A dinâmica da pesquisa está associada à freqüência das alterações verificadas entre cada período de visita e, portanto, requer um esforço permanente de atualização do sistema implantado.

Onde a pesquisa está sendo efetuada com maior rigor, ou seja, em Bocaiúva e Jacutinga, o número de informações é de, respectivamente, 17 para óbitos e 19 para nascimentos e 14 para óbitos e 12 para nascimentos. Em muitos casos o informante de nascimento é o mesmo de óbitos, mas onde é possível ter um para cada setor é preferível, por garantir maior independência de coleta. Para a coleta de estatísticas de nascimentos em quase todos os casos o informante é mulher e professora, enquanto para as de óbitos, encontram-se alguns homens, em geral moradores locais, que têm responsabilidade sobre o cemitério ou capela. Não existem preferências claras por idade, embora teoricamente as pessoas mais jovens tenham melhores condições de compreender e participar do trabalho.

Em Jacutinga, em função da forma concreta de organização daquela região de colonização européia, existe em cada povoado uma sociedade ou associação que congrega os moradores para fins de atividades religiosas e recreativas, cabendo ao presidente da sociedade, que é eleito anualmente, zelar pela integridade de bens comunitários como a igreja e o cemitério. Nesta situação, o informante natural de estatística é o próprio presidente, atualizando-se periodicamente o mesmo conforme a renovação anual. Em Bocaiúva alguns povoados também apresentam uma pessoa com atribuições parecidas, embora sem o grau de organização encontrado em Jacutinga.

A localização do povoado é o problema mais sério com que a pesquisa se depara em um município. Freqüentemente existem povoados não assinalados nos mapas e mesmo os agentes de coleta do IBGE desconhecem diversos. Em cada viagem inevitavelmente acabávamos por identificar um povoado não coberto e assim a relação do número de informantes é crescente. Parece que um dos passos iniciais para o tipo de coleta em experiência é desenvolver um esforço concentrado para o levantamento de todos os povoados existentes em um município selecionado. Isto posto, a seleção e a colaboração dos informantes não apresenta grandes dificuldades, excetuando-se o fato de não ser desejável que a colaboração se dê em níveis voluntários.

2.4 Procedimento de coleta

As visitas de nossa equipe de pesquisa aos municípios selecionados vêm ocorrendo, em média, duas vezes por ano, já tendo sido realizadas 6 visitas a Bocaiúva, 5 visitas a Jacutinga e 3 visitas a Bacabal.

Na parte da coleta de estatísticas de óbito, que foi iniciada em primeiro lugar, já contamos com dados para as três últimas viagens nos municípios de Bocaiúva e Jacutinga e na parte de nascimentos dispomos de dados desde a última viagem. O procedimento e os instrumentos de coleta diferem nos dois casos. Para óbitos utilizamos um formulário com perguntas sobre nome, sexo, idade, data do falecimento e local de moradia do indivíduo sepultado em cemitério pertencente ao povoado. Para o caso de crianças ou óbitos fetais perguntamos também o nome da mãe, visando a facilitar confrontos caso-a-caso com o Registro Civil ou outra qualquer fonte, já que em muitos casos a criança de poucos dias ou o óbito fetal ainda não tem nome. Atualmente estamos substituindo o formulário geral por fichas individuais, basicamente com as mesmas informações, mas que permitem simplificar o manuseio em escritório. O informante é instruído no sentido de coletar as informações para todos os sepultamentos realizados no povoado, independente de a pessoa ser ou não residente no mesmo. Em Bocaiúva, 2 vilas de distrito, por não possuírem cemitério com controle, também contam com informantes que cumprem a mesma instrução. Em Jacutinga os 2 distritos também contam com informantes, inclusive a sede que, no caso, é o secretário da Prefeitura que fornece a licença para o sepultamento.

Dispondo-se da cobertura de todos os povoados do município e tendo-se a informação por local de residência, pode-se afirmar que, teoricamente, estão sujeitos à coleta todos os óbitos de pessoas residentes no município e nele sepultadas. Entretanto, existem ainda duas formas de evasão de coleta de óbitos, sendo a primeira representada por aquelas pessoas residentes no município selecionado que vão ser sepultados em outros municípios e a segunda por sepultamentos ignorados (na maioria das vezes de crianças ou óbitos fetais) ou feitos no quintal das casas em áreas rurais ou em locais afastados. No primeiro caso a informação somente poderá ser recuperada quando o mesmo tipo de cobertura for estendido aos municípios vizinhos e, no segundo caso, é viável acreditar-se que a qualidade da coleta tende a melhorar desde que as condições se tornem favoráveis aos informantes, e então a cobertura desses casos individuais pode ser mais freqüente.

Para a coleta de estatísticas de nascimento adotou-se um procedimento específico que consiste em instruir o informante para preenchimento de uma ficha de gestante, que oferece a vantagem de "segurar"

a informação desde o momento em que a mulher estiver no período de gestação, possibilitando, assim, "amarrar" o resultado da gravidez. O informante deve coletar as informações em 3 etapas, ou seja, durante a gravidez, após o parto e 30 dias após o parto, o que, além de aumentar as probabilidades de coletar o nascimento, fornece também informações adicionais sobre mortalidade fetal e mortalidade neonatal, que podem ser comparadas e acrescentadas às fornecidas pelo informante de óbitos.

O sistema de informantes de nascimento é mais recente e tende a apresentar resultados mais fidedignos, a nível de povoado. Entre as informações fornecidas está uma sobre local de nascimentos, que permite avaliar a proporção de partos ocorridos em hospital e domicílio.

Paralelamente aos dados obtidos pelos informantes, o sistema pressupõe a coleta de dados dos cartórios do Registro Civil e dos hospitais existentes nos municípios sobre nascimentos e óbitos. Essas informações são confrontadas caso-a-caso com as dos informantes e permitem estabelecer os níveis de óbitos e nascimentos e estimar taxas de sub-registro por grupos de idade e sexo. Desta forma, o sistema usando as 3 fontes (informantes, cartório e hospital) dispõe de todas as informações alternativas possíveis em um município, e procedendo a um criterioso controle das informações de campo e a uma sistematizada organização dos resultados obtidos para cada período é capaz de combinar os dados através de *match* entre as fontes e estabelecer os dados estatísticos completos para o estudo da mortalidade e da natalidade em um município. O cotejo entre as fontes também permite visualizar as características da difusão dos serviços hospitalares e da influência do registro entre os povoados classificados por níveis de distância das sedes dos municípios, o que é bom indicador sobre os graus de contato e isolamento das populações rurais e sua exposição relativa aos sistemas tradicionais de estatísticas vitais.

Os procedimentos específicos de coleta empregados em campo são complementados na Divisão de Desenvolvimento Metodológico e Projetos Especiais do DESPO com uma organização sistemática de apuração e análise à base de cotejo e contagem das fichas procedentes das fontes diversas, sua avaliação, comparação e posterior ordenação em arquivo.

2.5 Perspectivas Futuras

Conforme analisaremos nas fases subseqüentes deste relatório, os resultados obtidos pela pesquisa de campo nos períodos considerados, a partir do 2.º semestre de 1974, foram excelentes, superando mesmo as expectativas iniciais. Durante essas etapas pôs-se em prática, com sucesso, um novo sistema de coleta de estatísticas vitais válido para áreas

predominantemente rurais e organizado de forma a gerar dados a nível de município. A pesquisa tem cumprido sua finalidade precípua, qual seja, a de demonstrar a exeqüibilidade do projeto em uma microexperiência. A extensão do projeto a níveis maiores (grupo de municípios, microrregião, unidade da Federação ou região) depende de sua avaliação não somente em função das perspectivas técnicas mas igualmente dos encargos a serem assumidos. O caráter complementar do sistema para áreas predominantemente rurais em relação a um sistema para as áreas urbanas (faltando definir o que neste contexto é "urbano", provavelmente cidades a partir de um certo tamanho não cobertas pelo sistema anterior) fornece nitidamente uma imagem de unidade dos sistemas de estatísticas vitais. A decisão sobre a utilização dos novos métodos pesquisados, a ampliação gradual ou imediata do sistema, a cobertura a um determinado nível de agregação, por amostra ou não, são questões que devem ser debatidas e decididas em função das necessidades e condições do IBGE.

Julgamos que a fase propriamente dita de pesquisa está quase terminando. Poucas viagens mais e a série de dados já estará bastante consistente. O aperfeiçoamento do sistema está, pois, agora, na dependência de sua própria expansão e adoção como um projeto normal do IBGE. As limitações demonstradas no decorrer da pesquisa são predominantemente devidas à carência de recursos e infra-estrutura com que a equipe do DESPO se defrontou. Na verdade, a experiência inicial dificilmente poderia ser realizada de outra maneira, em função de seu próprio caráter original e exploratório. A encampação oficial do projeto pelo IBGE estabelecerá as condições para que a plena utilização da infra-estrutura da Instituição seja mobilizada, principalmente a rede de coleta. O obstáculo principal enfrentado pela pesquisa foi, como já assinalamos, o fato de os informantes não serem remunerados. Para que o projeto obtenha um êxito capaz de se manter regular e constantemente, é necessário estudar formas de pagamento dos informantes e concomitantemente promover uma organização formal de treinamento em cada município que faça parte do projeto. Toda uma complexa interação de fases e decisões deverá ser criada para o desenvolvimento do projeto e tudo isso deverá ser levado em conta na avaliação de sua oportunidade. Períodos rigorosamente estabelecidos de coleta, de preferência trimestralmente, à maneira do Registro Civil, deverão ser fixados e conseqüentemente contabilizados. No DESPO será necessário atribuir a uma equipe técnica a função de avaliar, analisar e divulgar periodicamente os dados estatísticos.

As perspectivas futuras para um novo sistema de estatísticas vitais no Brasil são promissoras do ponto de vista técnico e metodológico, cabendo apenas definir sua viabilidade do ponto de vista financeiro e organizacional.

3. A COLETA DE ESTATÍSTICAS DE NASCIMENTOS: RESULTADOS E ANÁLISES

A coleta dos dados de nascimentos nas áreas de pesquisa obedeceu, como já foi descrito, a uma combinação entre o sistema experimental de informantes, o hospital e o cartório de Registro Civil.

Para a informante de nascimentos, que na maioria das vezes são professoras, foi criado um instrumento de coleta adequado à obtenção de dados sobre nascimentos ocorridos em áreas predominantemente rurais. Tal instrumento foi a ficha de gestante (*fac-simile* em anexo) cuja finalidade é acompanhar o fato vital desde o período de gestação, procurando-se, assim, tornar maior a probabilidade de obtenção da informação. A ficha, como já foi assinalado em capítulo anterior, comporta três etapas ou períodos de preenchimento, quais sejam: antes do parto (gestação), imediatamente após o parto e 30 dias após o parto. Inclui informações demográficas essenciais, como idade da mãe, residência habitual, resultado da gravidez e local de nascimento, bem como as informações referentes aos nomes e datas. É de especial interesse o fato de possibilitar o fornecimento de informações até 30 dias após o parto, o que contribui para enriquecer o nível dos dados sobre mortalidade infantil.

A ficha foi distribuída para todos os informantes designados para a coleta de nascimentos, ou seja, a nível de cada povoado encontrado e identificado nos três municípios.

A distribuição das fichas deu-se em junho de 1975 (Bocaiúva), agosto de 1975 (Jacutinga) e outubro de 1975 (Bacabal) e seus primeiros resultados foram coletados pela nossa equipe nas viagens subseqüentes, já no ano de 1976. Atualmente dispõe-se de dados relativos somente a este primeiro período e os resultados, do ponto de vista do preenchimento, foram os que inicialmente mais chamaram atenção dos técnicos. Com efeito, levando-se em conta o nível cultural de grande parte dos informantes e a precariedade do treinamento e da motivação dos mesmos, a ficha de gestante pode ser considerada razoavelmente complexa e, por conseguinte, de preenchimento mais difícil. Assim sendo, estabeleceu-se um "plano de crítica", visando a identificar os tipos mais freqüentes de erro e refletir sobre a conveniência de alterar-se ou não certos quesitos. Teoricamente, constatou-se que poderiam haver quatro grandes grupos de erros de preenchimento:

— erros por omissão

— erros por incoerência de datas

— erros por múltiplo preenchimento

— erros por estar fora do período ou local de estudo.

Esses quatro grandes grupos de erros foram considerados a partir de 3 categorias não excludentes que definem um critério prático:

— erros que não afetam a qualidade da informação

— erros que diminuem a qualidade da informação

— erros que levam à anulação da ficha.

Para fins práticos, considerou-se que algumas omissões não significantes (nome da criança, local de nascimento, por exemplo) e pequenas incoerências de datas, embora pudessem alterar a qualidade das informações, não levavam necessariamente à anulação da ficha. Os casos em que a anulação foi feita, para Bocaiúva, apresentam o seguinte quadro:

Omissão	2 (1,01%)
Preenchimento múltiplo	1 (0,51%)
Fora do período ou local	17 (8,63%)
TOTAL	20 (10,15%)

Deve-se ressaltar que as informações fora do período da pesquisa ou do local (município) não constituem informações erradas e sim dados que estão fora do âmbito da pesquisa.

O fato de somente 10% das fichas terem sido anuladas no município de Bocaiúva quando da primeira coleta é altamente encorajador, principalmente levando-se em conta as limitações já enumeradas. Evidentemente, outras informações estão sujeitas a imprecisões, como as diversas datas constantes da ficha e, nesses casos, para que haja uma melhoria segura da qualidade das informações faz-se necessária a volta ao campo para verificação dos casos individuais.

De acordo com o sistema de coleta proposto, o passo seguinte consistiu em realizar o confronto ou *match,* caso a caso, entre as fichas de nascidos vivos do hospital e os registros de nascimentos dos cartórios. Essa fase também interessou secundariamente ao próprio controle das informações fornecidas pelas fichas, visto a possibilidade de checar os diversos itens (datas, nomes, locais, etc.) quando as informações são identificadas como referentes à mesma pessoa.

Os problemas técnicos surgidos na operação de confronto são, em geral, numerosos, residindo os principais nos erros quanto às datas de nascimento, nome da mãe e local de nascimento. O primeiro tipo de erro é o mais comum, sendo aceitável imediatamente uma diferença de dias e caso a diferença seja de meses, por exemplo, torna-se necessário controlar rigorosamente o confronto entre as outras informações existentes, para que seja possível a decisão sobre a identidade dos eventos.

Uma dificuldade adicional que se apresentou ao fazer-se o confronto caso a caso foi o fato de as informações dos cartórios e hospitais não estarem em forma de fichas individuais, mas em formulários (ou listas) com um evento por linha, o que torna mais vagarosa e complexa a operação de confronto. Após essa experiência, decidiu-se utilizar fichas individuais para todas as fontes, o que permite um agrupamento rápido e eficiente das informações.

O número de casos em que o confronto foi positivo será exposto na fase seguinte que apresenta e analisa os resultados obtidos.

3.1 Apresentação Analítica dos Resultados

Após a primeira coleta de fichas de gestante para os três municípios de estudo, obteve-se resultados que exprimem um quadro bastante geral para Bocaiúva e Jacutinga, enquanto que para Bacabal (MA) os resultados cobrem apenas uma determinada área do município, onde foi fisicamente possível implantar a pesquisa. Os resultados para os dois primeiros municípios são, levando-se em conta tratar-se da primeira coleta, inegavelmente bons, enquanto para Bacabal o sistema ao menos mostrou-se viável onde foi implantado, pondo por terra algumas especulações sobre a inexeqüibilidade do mesmo nas regiões econômica e culturalmente atrasadas do Nordeste.

Apresentaremos, em seguida, os resultados obtidos para o município de Bocaiúva no período que vai de 19-6-75 a 18-2-76, ou seja, aproximadamente 8 meses.

Na tabela 2 encontramos os dados globais coletados para as 3 fontes de dados sobre nascimentos ocorridos no período acima. Pode-se ver que o hospital apresentou a maior cobertura geral, tanto para a sede como para o interior do município. O hospital informa também alguns eventos ocorridos fora de suas dependências, pelo fato de suas visitadoras sanitárias realizarem atendimentos e visitas nos domicílios. No caso do cartório observa-se uma tendência maior de registro para os nascimentos ocorridos em hospital, o que serve de apoio à tese de que o ato de registro está relacionado, em boa margem, ao grau de acesso que as populações têm aos serviços institucionalizados, grau de acesso este que está condicionado por vários fatores (econômicos, culturais, geográficos, etc.). O sistema de informantes apresentou uma cobertura superior ao cartório, mesmo sendo utilizado somente fora da sede municipal. Quase 2/3 dos dados fornecidos pelos informantes referem-se a nascimentos ocorridos em domicílios, invertendo assim o que se observa nas outras fontes.

Uma análise mais detalhada de tabela 1 deixa patente que a fonte oficial de estatísticas vitais, ou seja, o cartório, é a que demonstrou menor grau de cobertura. Trata-se de um resultado altamente enco-

rajador para quem, como nós, propõe métodos alternativos e/ou complementares de coleta de estatísticas vitais. Em segundo lugar, sabendo-se que, para 1970, a população da sede era quase 3 vezes inferior à população do resto do município, o fato de que os nascimentos coletados no hospital (262 fora da sede e 196 na sede) e no cartório (86 fora da sede e 75 na sede) apresentem uma diferença bastante diminuta entre os nascimentos ocorridos nas duas áreas, evidencia que tanto o hospital como principalmente o cartório dispõem de uma possibilidade extremamente limitada de cobrir a grande maioria dos nascimentos ocorridos fora da sede municipal. É justamente neste ponto que o sistema de informantes pode oferecer a sua grande contribuição, alcançando significativamente os nascimentos ocorridos em domicílios no interior dos municípios.

TABELA 2

NASCIMENTOS COLETADOS NAS TRÊS FONTES, POR SEXO, SEGUNDO A SITUAÇÃO DO DOMICÍLIO E O LOCAL DO NASCIMENTO
BOCAIÚVA — 19-06-75 A 18-02-76

SITUAÇÃO DE DOMICÍLIO	LOCAL DE NASCIMENTO	NASCIMENTOS, POR FONTES DE COLETA								
		Cartório(1)			Hospital			Informante(2)		
		Masc.	Fem.	Total	Masc.	Fem.	Total	Masc.	Fem.	Total
Sede Municipal	Domicílio	3	3	6	14	10	24	—	—	—
	Hospital	32	37	69	83	89	172	—	—	—
	Total	35	40	75	97	99	196	—	—	—
Fora da Sede Municipal	Domicílio	17	21	38	1	1	2	—	—	114
	Hospital	29	19	48	133	127	260	—	—	60
	Ignorado	—	—	—	—	—	—	—	—	2
	Total	46	40	86	134	128	262	—	—	176
TOTAL	Domicílio	20	24	44	15	11	26	—	—	114
	Hospital	61	56	117	216	216	432	—	—	60
	Ignorado	—	—	—	—	—	—	—	—	2
	Total	81	80	161	231	227	458	—	—	176

FONTE — Cartório de Registro Civil, hospitais e fichas de gestantes preenchidas pelos informantes

(1) Faltou a coleta para o Distrito de Olhos D'Água.

(2) Faltou a coleta para 4 povoados e as informações referentes aos meses de janeiro e fevereiro (até o dia 18) de 1976 para o povoado de Engenheiro Dolabela.

A partir desta última constatação, resolvemos buscar, nas fichas coletadas junto aos informantes, outras indicações que pudessem contribuir para uma melhor compreensão das características das gestantes residentes fora da sede. A idéia principal era procurar fatos que expli-

cassem a maior incidência de partos em domicílios (nas vilas e áreas rurais) mesmo tendo em vista a existência de hospital na sede do município. É evidente que a nível macroestrutural os determinantes econômicos e culturais são os mais importantes. Mas também outras características, que estão associadas aos níveis determinantes mais gerais, podem ter um papel específico na decisão sobre o local de nascimento. Uma das hipóteses, muito corrente em certos setores das Ciências Sociais, é a de que as novas gerações, por força de maior sensibilidade às mudanças recentes, são agentes mais eficazes nos processos de assimilação dos novos costumes e dos novos valores. Transpondo-se a formulação para o nosso caso específico, a hipótese é de que as mães mais jovens teriam maior propensão a substituir o tradicional parto em domicílio pelo hospital. Entretanto, com base nos dados constantes da tabela 3, não se observou a esperada correlação entre a idade da gestante e o local do parto, o que corresponde a admitir que outras explicações devem ser buscadas para o fato.

TABELA 3

NASCIMENTOS OCORRIDOS, POR LOCAL, SEGUNDO OS GRUPOS DE IDADE DAS GESTANTES
BOCAIÚVA — 19-06-75 A 18-02-76

GRUPOS DE IDADE DAS GESTANTES	NASCIMENTOS, SEGUNDO O LOCAL			
	Total	Hospital	Domicílio	Ignorado
TOTAL	176	60 (34%)	114	3
15⊢20	20	10 (50%)	10	—
20⊢25	36	10 (29%)	25	1
25⊢30	35	10 (29%)	24	1
30⊢35	22	7 (32%)	15	—
35⊢40	25	11 (44%)	14	—
40⊢45	13	6 (46%)	7	—
Idade Ignorada	25	6 (24%)	19	—

FONTE — Fichas de gestantes preenchidas pelos informantes.
NOTA — Resultados obtidos fora da sede municipal.

Uma idéia já anteriormente desenvolvida é que quanto maior o isolamento a que estão submetidas as populações rurais (incluindo as vilas e povoados) mais tendem a demonstrar apego aos seus sistemas tradicionais. O conceito de isolamento deve ser tomado inicialmente no sentido cultural e econômico, mas também pode ser aplicado ao ponto de vista físico de distância. Poderíamos dizer que quanto mais difícil

o acesso à sede municipal menos a área em questão é atingida pelos sistemas institucionalizados em questão (hospitais e cartórios). Tomamos a distância física como uma primeira aproximação de dificuldade de acesso à sede.

Na tabela 4 agrupamos os povoados do município de Bocaiúva em classes de distância (km) da sede e verificamos para cada classe o número de partos em hospitais e em domicílio. Os dados demonstram inequivocamente que existe correlação positiva entre a distância e o local do parto, na medida em que quanto maior é a distância mais ocorrem partos em domicílios, chegando mesmo, na última classe, a atingir quase a totalidade dos partos.

Por conseguinte, existem razões objetivas que determinam a não cobertura dos eventos pelos sistemas tradicionais de estatísticas vitais, tanto o oficial (cartório), como o imediatamente disponível (hospital). Tais razões podem ser atribuídas às formas de organização específica que predominam em áreas rurais ou afastadas das sedes municipais e também a todo um conjunto de fatores econômicos e sócio-culturais que caracterizam as diversas regiões do País. A atual pesquisa de novos métodos de coleta de estatísticas vitais é um passo importante no sentido de apontar as limitações dos sistemas usuais e propor o conjunto de sistemas capazes de fornecer as estatísticas desejadas.

TABELA 4

NASCIMENTOS OCORRIDOS, POR LOCAL, SEGUNDO AS CLASSES DE DISTÂNCIA DA SEDE
BOCAIÚVA — 19-06-75 A 18-02-76

CLASSES DE DISTÂNCIA DA SEDE (km)	NASCIMENTOS, SEGUNDO O LOCAL			
	Total	Hospital	Domicílio	Ignorado
TOTAL	176	60	114	2
Até 20	26 (100%)	16 (62%)	10 (38%)	—
20|–40	67 (100%)	26 (40%)	39 (60%)	2
40|–60	62 (100%)	17 (27%)	45 (73%)	—
60 e mais	21 (100%)	1 (5%)	20 (95%)	—

FONTE — Fichas de gestantes preenchidas pelos informantes.
NOTA — Resultados obtidos fora da sede municipal.

Por último, as informações contidas nas fichas de gestantes foram comparadas caso a caso com as dos hospitais e dos cartórios e, mesmo considerando-se as limitações devidas ao fato de se dispor somente do período referente à experiência, os resultados do *match* foram bons. À

parte algumas pequenas diferenças quanto às características informadas, conforme assinalado na parte introdutória, o *match* veio a confirmar o grau de confiança atribuído pela equipe de pesquisa às informações fornecidas pelo sistema de informantes de estatísticas de nascimento.

Na tabela 5 apresentamos o resultado do *match* para as 3 fontes, conforme apurado para os nascidos vivos de mães residentes fora da sede municipal, ou seja, nas vilas e áreas rurais. Observe-se que, de um total de 448 nascimentos coletados, cerca de 377 (136 + 199 + 42) são nascimentos coletados por apenas uma das 3 fontes, enquanto 71 (5 + 8 + 27 + 31) são casos em que ocorre *match*, ou seja, mais de uma fonte coleta o mesmo nascimento. O total de nascimentos efetivamente coletados para o período foi, por conseguinte, de 71 + 377, ou seja, de 448 nascimentos. Na tabela 2, anteriormente comentada, tínhamos para o total de nascidos vivos de mães residentes fora da sede municipal cerca de 524 casos perfazendo uma diferença de 76 casos de múltipla observação. Seriam, portanto, os 71 duplamente observados mais 5 que foram triplamente observados, ou seja, observados nas 3 fontes.

TABELA 5

CONFRONTO ENTRE OS NASCIMENTOS, SEGUNDO AS TRÊS FONTES DE COLETA
BOCAIÚVA — 19-06-75 A 18-02-76

FONTE 2: FICHA DE GESTANTE (INFORMANTE DE NASCIMENTO)	FONTE 3: CARTÓRIO	FONTE 1: HOSPITAL		
		Observado	Não Observado	Total
Observado	Observado	5	8	13
	Não observado	27	136	163
	Total	32	144	176
Não observado	Observado	31	42	73
	Não observado	199	—	199
	Total	230	42	272
Total	Observado	36	50	86
	Não observado	226	136	362
	Total	262	286	448

FONTE — Cartórios do Registro Civil, hospitais e fichas de gestantes preenchidas pelos informantes.
NOTA — Resultados obtidos somente nas vilas e áreas rurais.

A experiência do *match* foi altamente proveitosa, servindo para aferir a qualidade das informações e, importante, demonstrando que um sistema de coleta que combine formas alternativas e complementares é de grande eficácia para uma considerável melhoria das estatísticas de nascimento.

4. SUGESTÕES PARA A IMPLANTAÇÃO DE UM SISTEMA DE ESTATÍSTICAS VITAIS NO BRASIL

A experiência adquirida na fase inicial do SEV (Sistema de Estatísticas Vitais) nos leva a fazer as seguintes sugestões:

1 — Que não seja implantado de uma vez só no País todo. Ao contrário, que se inicie com uns poucos lugares, nos quais a implantação seria cuidadosa e controlada. Pouco a pouco, então, seriam agregadas novas áreas. A implantação imediata em todo o País acarretaria, no nosso entender, um baixo nível na qualidade do produto final. Seria muito difícil controlar e manter uma boa qualidade dos dados se já de início, sem uma grande estrutura montada para esse fim, recebermos uma quantidade muito grande de material para ser acompanhado, controlado e processado. Parece-nos mais prudente iniciarmos com umas poucas localidades, nas quais seria despendido um grande esforço no sentido de não somente manter uma alta qualidade da informação como também nos permitiria introduzir aperfeiçoamentos ainda nessa fase.

No caso da coleta de dados sobre nascimentos ocorridos nos hospitais, parece-nos que o mais aconselhável é iniciarmos com uma cidade de tamanho médio que conte com uma rede hospitalar bastante boa. Este foi o motivo pelo qual, na fase experimental, os técnicos do DESPO mantiveram contatos com as Secretarias de Saúde, diretamente, e não com o Ministério da Saúde. Tais contatos foram feitos com as Secretarias de Saúde do Rio Grande do Sul, Santa Catarina, Paraná e Minas Gerais.

2 — Sugerimos a assinatura de convênios com algumas dessas Secretarias de Saúde, com a finalidade de implantar o Sistema de Estatísticas Vitais em alguns desses estados. Os contatos feitos até o momento foram sempre informais e entre técnicos do DESPO e daquelas secretarias. E, apesar dos excelentes resultados conseguidos ao nível de planejamento, a execução de qualquer programa torna-se extremamente difícil, se não impossível, sem que este seja oficializado. Em quase todas as ocasiões, os contatos que os técnicos do DESPO tiveram com os técnicos das secretarias tiveram que ser sistematicamente refeitos, em virtude de constantes mudanças de pessoal nas secretarias. Para que o trabalho conjunto não sofra solução de continuidade torna-se necessária a assinatura de convênios específicos. Parece-nos que ao nível de

contatos informais, o projeto já foi tão longe quanto seria aconselhável. Daqui por diante quase nada se pode fazer sem a oficialização desses contatos.

3 — Gostaríamos de sugerir que, paralelamente ao sistema de produção de dados, fosse criado, como parte inseparável dele, um mecanismo rigoroso de controle de qualidade que abrangesse todas as fases, principalmente nas primeiras. Não se trata somente de garantir que o dado produzido na fonte chegue quase intato na fase de divulgação, mas que a fonte em si seja sistematicamente objeto de avaliação crítica, uma das quais seria o confronto caso a caso (*match*) com outra fonte independente, como, por exemplo, as pesquisas domiciliares e os censos de população. Daí uma vantagem de se começar a implantação do SEV em municípios que façam parte da PNAD. Isto é extremamente importante, uma vez que os sistemas contínuos, entre os quais estão as estatísticas contínuas, tendem inexoravelmente a deteriorar a qualidade do produto caso não haja um esforço constante no sentido de evitar que tal ocorra. Isto é necessário para, no mínimo, manter a qualidade do sistema. O ideal, no entanto, seria ir um pouco além; que este mecanismo fosse não só de controle de qualidade, como de aperfeiçoamento. A não existência de tal mecanismo, acreditamos, fatalmente levará a degradação progressiva dos dados.

É preciso que, periodicamente, se conheça a taxa de evasão do sistema.

4 — O censo de 1980 poderia ser utilizado para estimar o grau de cobertura do registro civil de óbitos, através da inclusão de uma pergunta sobre óbitos ocorridos nos 3 meses anteriores ao censo. Essa informação seria então cotejada, caso a caso, com os óbitos registrados, obtendo-se assim uma estimativa do número real de óbitos ocorridos. Dado sua importância, parece-nos que este assunto, juntamente com o problema de estimar o grau de cobertura do próprio censo (não se sabe no Brasil a taxa de cobertura de nenhum dos censos e, conseqüentemente, o erro nas taxas de crescimento demográfico), merecem ao menos serem discutidos.

O desenvolvimento da análise demográfica é inegável, entretanto, no que se refere às variáveis básicas, taxas brutas de natalidade e mortalidade, taxas de mortalidade infantil e taxa de crescimento, a qualidade da informação (baixa, por sinal) pouco melhorou nos últimos 40 anos. Sem conhecermos as taxas de cobertura dos censos de 1960 e 1970 não é possível nem mesmo ter certeza de que o pequeno declínio aparente da taxa de crescimento seja real. O conhecimento dessas variáveis com boas estimativas de seus erros é necessário para uma demografia quantitativa com fundações sólidas.

O SEV adotou, com sucesso, alguns métodos que acreditamos nunca terem sido utilizados antes. Um deles é a sistemática de iniciar a coleta de informações sobre nascimentos localizando primeiramente todas as gestantes da localidade e acompanhá-las até 30 dias após o parto. Esse procedimento evita uma série de inconvenientes dos outros métodos, como, por exemplo, a sistemática omissão em declarar crianças que faleceram logo após o parto (em algumas regiões, se a criança morre antes de ser batizada, as pessoas tendem a ignorar sua existência), óbitos fetais, etc. Não sendo considerados fatos importantes, são sistematicamente omitidos quando se coletam informações sobre nascimentos. Há vários outros exemplos de métodos inovados pelo SEV, mas como constam de relatórios específicos não faremos aqui referência a eles.

Parece-nos que o SEV, apesar de ainda não implantado (produzindo dados), já produziu resultados no que concerne a métodos de pesquisa, sendo que sua experiência poderia ser aproveitada em outros países, em particular, da América Latina, pois em vários lugares a situação deve ser semelhante.

Sugerimos, finalmente, que qualquer sistema que se adote para a obtenção de estatísticas vitais seja tal que o fornecimento dos dados sobre o nascimento independa da vontade dos pais e parentes da criança. Achamos que a responsabilidade da declaração do nascimento para fins estatísticos deve caber a um profissional. Caso contrário, somos bastante céticos quanto à grande cobertura do sistema.

ANEXOS

MUNICÍPIO DE BOCAIÚVA (SOMENTE VILAS E ÁREAS RURAIS)

Confronto entre as três fontes de dados

FICHA DE GESTANTE — N I

INFORMAÇÕES OBTIDAS ANTES DO PARTO

NOME DA GESTANTE — DATA PROVÁVEL DO PARTO mes / ano

RESIDÊNCIA HABITUAL (rua, número, cidade ou povoado, município)

IDADE — DATA EM QUE ESTAS INFORMAÇÕES FORAM OBTIDAS dia / mes / ano

INFORMAÇÕES OBTIDAS LOGO APÓS O PARTO

DATA DO PARTO / / dia mes ano

RESULTADO DA GRAVIDEZ

☐ nascido vivo ☐ nascido morto ☐ aborto

LOCAL DO NASCIMENTO

☐ hospital ☐ domicilio ☐ outro

NOME DA CRIANÇA

SEXO

☐ masculino ☐ feminino

DATA EM QUE ESTAS INFORMAÇÕES FORAM OBTIDAS dia / mes / ano

INFORMAÇÕES OBTIDAS 30 DIAS APÓS O PARTO (sòmente para crianças nascidas vivas)

☐ A CRIANÇA AINDA ESTAVA VIVA 30 DIAS APÓS O PARTO

☐ A CRIANÇA FALECEU ANTES DE COMPLETAR 30 DIAS DE VIDA

DATA DO FALECIMENTO / / dia mes ano

DATA EM QUE ESTAS INFORMAÇÕES FORAM OBTIDAS dia / mes / ano

OUTRAS INFORMAÇÕES

Nº de ordem

NOME DO INFORMANTE:

LUGAR DE RESIDENCIA:

FICHA DE SEPULTAMENTOS OCORRIDOS

NOME DO FALECIDO:

SEXO ☐ masculino ☐ feminino

IDADE:

DATA DE FALECIMENTO dia / mes / ano.

RESIDÊNCIA HABITUAL < povoado, vila ou cidade / municipio

O I

DATA DA DISTRIBUIÇÃO

DATA DA COLETA

NOME DA MÃE (somente para menores de 18 anos)

NÚMERO DE ORDEM

OUTRAS INFORMAÇÕES

NOME DO INFORMANTE

LUGAR DE RESIDÊNCIA

ESTADO DO PARANÁ
SECRETARIA DE ESTADO DA SAÚDE E DO BEM ESTAR SOCIAL

1ª VIA

ATESTADO DE NASCIMENTO VIVO

Nº _____ A

CRIANÇA

1 MUNICÍPIO

2 NASCIDO VIVO (nome) | 3 DATA NASC. ___/___/___

4 LOCAL DO NASCIMENTO: 1 - Hospital ☐ 2 - Outro ☐

5 SEXO: 1 - Masc. ☐ 2 - Fem. ☐

6 PESO AO NASCER: _____ Gramas

7 GRAVIDEZ: 1 - Simples ☐ 2 - Gemeos ☐ 3 - Trigem. ☐ 4 - + de 3 ☐

8 TEMPO DE GESTAÇÃO (semanas)

9 TIPO DO PARTO: 1 - Expontâneo ☐ 2 - Operatório ☐ 3 - Forceps ☐

10 PARTO ATENDIDO POR: 1 - Médico ☐ 2 - Obstetriz ☐ 3 - Enfermeiro ☐ 4 - Outro ☐

MÃE

11 MÃE (nome completo) | 12 DATA NASC. ___/___/___

13 IDADE NA OCASIÃO DO PARTO | 14 NATURALIDADE (Unidade da Federação ou País Estrangeiro)

15 RESIDÊNCIA HABITUAL (rua, número, bairro - vila ou povoado)

16 NA SEDE MUNICIPAL: 1 - Sim ☐ 2 - Não ☐ | 17 MUNICÍPIO | 18 UNID. FED.

PAI

19 PAI (nome completo) | 20 IDADE | 21 NATURALIDADE

CARTÓRIO

22 CARTÓRIO EM QUE DEVERÁ SER REGISTRADO ESTE NASC. (endereço)

ATESTANTE

23 ATESTANTE

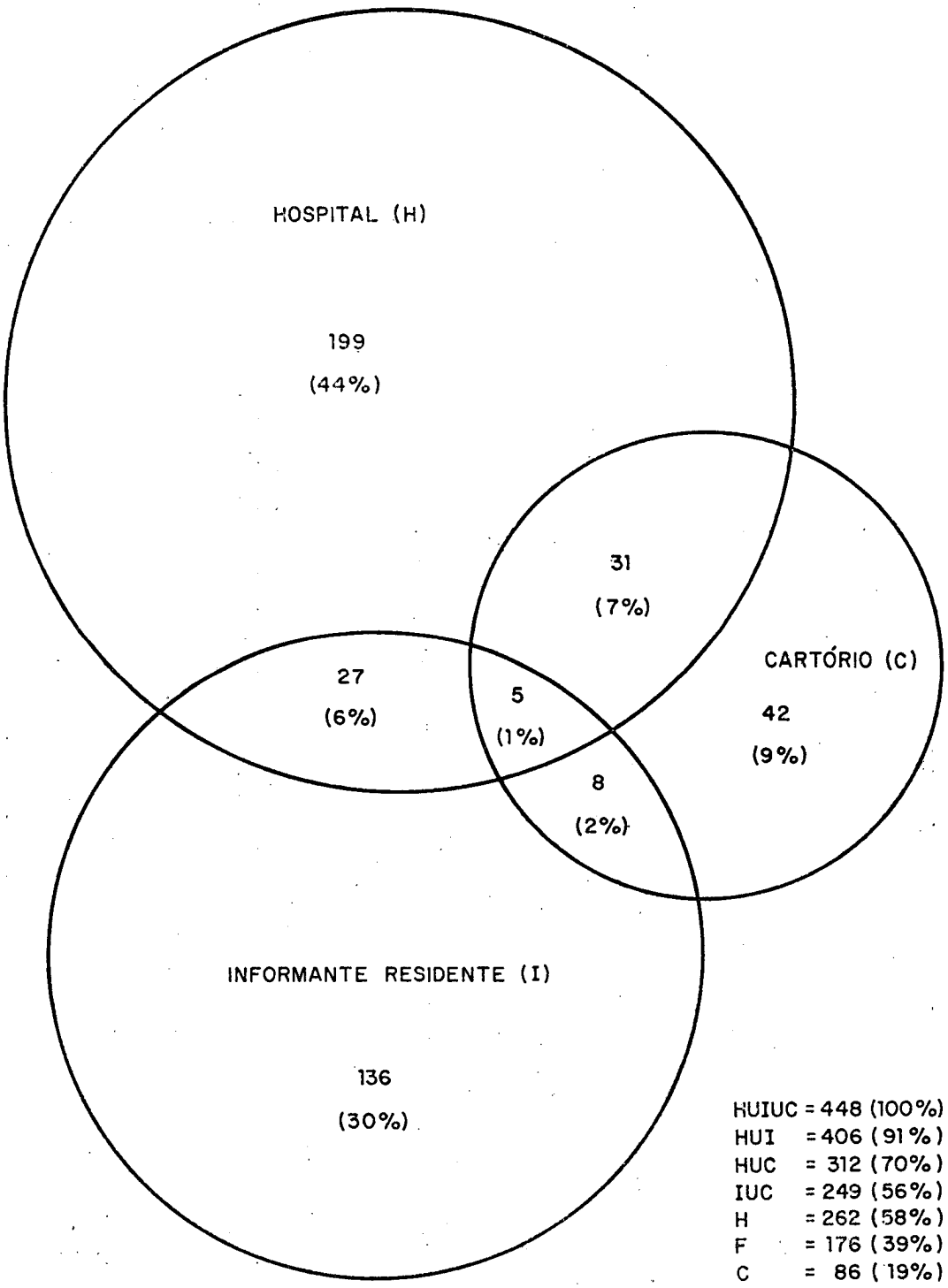
24 ENDEREÇO COMPLETO

25 TELEFONE | 26 DATA ___/___/___ | 27 ASSINATURA

CARIMBO DO HOSPITAL OU INSTITUIÇÃO

MUNICÍPIO DE BOCAIÚVA (SOMENTE VILAS E ÁREAS RURAIS)

Confronto entre as três fontes de dados



HUIUC = 448 (100%)
HUI = 406 (91%)
HUC = 312 (70%)
IUC = 249 (56%)
H = 262 (58%)
F = 176 (39%)
C = 86 (19%)

ABSTRACT

Utilizing data from a field test carried out in the município of Bocaiuva, the problem of obtaining reliable statistics on births and deaths is examined comparing three different sources of data (civil registration, hospitals, and local informants).

Data on births in rural areas was obtained from a "pregnancy file" producing very promissing results.

The proportion of births ocurring in hospitals and the proportion of births registered according to the National Household Survey of 1973 (PNAD 73) are presented and for the large cities, the collection of data on births directly from the hospital is proposed.

It is also suggested that every five years a survey to estimate the coverage of the data on births and deaths should be carried ont.

A MORTALIDADE NAS REGIÕES METROPOLITANAS*

Celso Cardoso da Silva Simões
Chefe da DIESM/DESPO

SUMÁRIO

RESUMO

As esperanças de vida obtidas permitiram verificar que a mortalidade é maior nas regiões metropolitanas menos desenvolvidas. Os resultados indicaram a existência de extremas diferenças de mortalidade para cada região.

Foi observado que existe um setor majoritário da população de baixos rendimentos nos quais as crianças estão expostas a riscos muito altos de mortalidade, em oposição àqueles segmentos da população de maiores rendimentos, onde as crianças estão expostas a menores riscos.

* Este estudo é um capítulo da tese *O Quadro da Mortalidade por Classes de Renda: Um Estudo dos Diferenciais nas Regiões Metropolitanas (Núcleo e Periferia)*, submetida e aprovada pelo Corpo Docente da Coordenação de Pós-Graduação de Engenharia da Universidade Federal do Rio de Janeiro, para obtenção do Grau de Mestre em Ciências (M.Sc.).

Infere-se também, dos resultados encontrados, que as disparidades na distribuição da renda monetária e no acesso à infra-estrutura urbana (sistemas adequados de água e esgoto) de bens e serviços, entre os diversos segmentos de população, são associados com a desigualdade entre os níveis de mortalidade encontrados para os estratos de mais baixa renda e os de alta renda.

1. INTRODUÇÃO

Este capítulo tem por objetivo determinar os níveis de mortalidade por estratos de renda nas regiões metropolitanas brasileiras, utilizando para isso a técnica de Sullivan [1].

Procura-se relacionar os níveis de mortalidade encontrados para cada região, com o acesso das populações a serviços de infra-estrutura urbana (água e esgoto), bem como com as condições de nutrição, referidas apenas às Regiões Metropolitanas de Recife e de Porto Alegre.

Os dados básicos utilizados foram tabulações especiais da amostra do censo demográfico de 1970 e dados do Estudo Nacional de Despesa Familiar (ENDEF-1974/75).

2. VIDA MÉDIA AO NASCER PARA O TOTAL DAS REGIÕES METROPOLITANAS

Em termos teóricos, a vantagem da utilização do número médio de anos ao nascer (l_o) como indicador do padrão de vida das populações das regiões metropolitanas brasileiras reside nos seguintes aspectos:

a) Por ser expresso em anos, é um conceito de fácil compreensão;

b) por sumarizarem as taxas de mortalidade de todas as idades, não estão afetadas pelas diferentes composições etárias da população, razão pela qual refletem as condições gerais de mortalidade prevalecentes, podendo ser utilizadas para fins de comparação entre as diferentes populações.

A tabela 1 apresenta as taxas de esperança de vida ao nascer para as 9 regiões metropolitanas brasileiras, segundo a condição de naturalidade de sua população, estimadas com base nos valores ${}_2q_0$, ${}_3q_0$ e ${}_5q_0$.

São bem conhecidas as disparidades regionais no Brasil, ligadas a fatores históricos e sócio-econômicos. A distribuição geográfica desigual

[1] Para detalhes sobre o desenvolvimento desta técnica, vide:

SULLIVAN, Jeremiah R. — Models for estimation of the probability of dying between birthand exact ages of early childhood. *Population Studies*. London, 26(1), 79-89, mar., 1972.

TABELA 1

VIDA MÉDIA AO NASCER, OBTIDA COM BASE NA MORTALIDADE DOS FILHOS MENORES DE 5 ANOS, POR CONDIÇÃO DE NATURALIDADE, SEGUNDO AS REGIÕES METROPOLITANAS

REGIÕES METROPOLITANAS	VIDA MÉDIA AO NASCER			
	Total	Naturais	Migrantes	Naturais/ Migrantes
Belém	55,06	55,74	53,69	1,04
Fortaleza	41,81	43,09	40,26	1,07
Recife	47,05	48,12	45,88	1,05
Salvador	48,20	48,84	47.37	1,03
Belo Horizonte	52,99	56,53	51,68	1,09
Rio de Janeiro	56,00	59,17	54,05	1,09
São Paulo	56,58	61,31	55,54	1,10
Curitiba	55,66	56,31	54,99	1,02
Porto Alegre	60,51	61,41	59,95	1,02
Conjunto Metropolitano	54,03	55,33	53,09	1,04

FONTE — Estimativas utilizando o Método de Brass-Sullivan, obtidas tomando por base: IBGE, Censo Demográfico de 1970, Tabulações Especiais, Rio de Janeiro.

dos fatores da produção leva a níveis muito diversificados de desenvolvimento econômico, o que é refletido nas amplas diferenças dos níveis de mortalidade encontrados. As regiões metropolitanas do Nordeste, por exemplo, mostram esperanças de vida substancialmente mais baixas em relação às regiões metropolitanas mais desenvolvidas do centro-sul do país. A associação entre nível de desenvolvimento sócio-econômico e níveis de mortalidade de cada região metropolitana é bastante claro, com um coeficiente de correlação de Spearman na ordem de 0,86 [2].

Dentro do próprio Nordeste encontram-se acentuadas desigualdades entre as regiões metropolitanas. Neste sentido, convém ressaltar a precária situação, em termos de esperança de vida, das crianças de Fortaleza, com valor na ordem de 41,81 anos, representando um nível de mortalidade só comparável ao existente nos países europeus por volta de 1870 [3].

Este nível contrasta com o verificado para as regiões metropolitanas do centro-sul, em especial Porto Alegre, com esperança de vida de aproximadamente 61 anos. Assim, as crianças desta região metropolitana vivem em média mais de 18 anos que as de Fortaleza.

[2] TUCCI NETO, E. — A ordenação das regiões metropolitanas brasileiras, segundo seu nível sócio-econômico.

[3] United Nations — *The Determinant and Consequences of Population Trends.* Departament of Economic and Social Affairs. New York, 1975.

Por outro lado, ao se considerar a condição de naturalidade das populações nas regiões metropolitanas, verifica-se que os filhos de migrantes estão expostos a uma mortalidade mais elevada do que os filhos dos naturais.

Este diferencial é acentuadamente maior somente nas mais importantes regiões metropolitanas de imigração — Rio de Janeiro e São Paulo — 9% e 10%, respectivamente, onde existe uma elevada proporção de migrantes oriundos do Nordeste, bem como na Região Metropolitana de Belo Horizonte.

Entretanto, quando se considera o conjunto das regiões metropolitanas, naturais e migrantes não diferem muito quanto ao nível de mortalidade. Desta forma, os resultados obtidos favorecem a hipótese de que as diferenças encontradas a nível de região metropolitana para migrantes e naturais não devem ser privilegiadamente explicadas por pretensas, ou não, diferenças de qualificação e/ou oportunidades entre eles, mas sim pelas condições históricas das formações econômicas e sociais a nível regional. Desta forma, no que respeita aos níveis de mortalidade alcançados em cada região, os determinantes não são encontrados na condição de migrante ou natural e sim no estágio de desenvolvimento econômico-social que determina os níveis. Assim, "migrantes e naturais são partes constituintes de uma só população que está submetida às mesmas leis econômicas e sociais de funcionamento do mercado.

A intensidade e as características destas leis são variáveis, conforme o nível histórico de desenvolvimento das relações capitalistas de produção. Inclusive, sob certas circunstâncias historicamente definidas, a manutenção das desigualdades regionais ou internas é, paradoxalmente, uma forma de promover a continuidade deste desenvolvimento" [4].

A título de ilustração (ver tabela 1), as regiões metropolitanas do centro-sul do país ostentam mortalidade mais baixa, tanto para migrantes como para naturais, o que nos leva a concluir que estas populações teriam condições de vida melhores que as das outras regiões.

Confronte-se os dados sobre mortalidade com alguns indicadores de nível de vida disponíveis (tabela 2), para cada região metropolitana.

Em relação às variáveis indicadoras do estado sanitário da população, observa-se que a existência de instalações sanitárias satisfatórias e de água encanada nos domicílios diminuiria, em muito, a possibilidade de ocorrência de doenças, tais como: a febre tifóide, paratifóide, disenteria e outras doenças infecciosas e parasitárias, particularmente perigosas durante a infância. Sabe-se que o armazenamento de água em depósitos possibilita a sua contaminação; da mesma forma, a água

[4] OLIVEIRA, L. A. P. — Aspectos econômicos das famílias de chefe migrante e natural. In: CASTRO, Mary et alii. *O quadro das famílias em domicílio de chefe migrante e natural*: Um estudo censitário dos diferenciais nas regiões metropolitanas brasileiras. Rio de Janeiro, MINTER/IBGE, 1977, 232 p. (mimeo).

TABELA 2

INDICADORES DE NÍVEL DE VIDA POR REGIÕES METROPOLITANAS — BRASIL — 1970

	INDICADORES	REGIÕES METROPOLITANAS								
		Belém	Fortaleza	Recife	Salvador	Belo Horizonte	Rio de Janeiro	São Paulo	Curitiba	Porto Alegre
					SANEAMENTO					
I)	Instalações Sanitárias em Domicílios (%)	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Rede Geral	8,06	2,39	12,95	9,91	34,17	38,57	58,12	21,50	22,70
	Fossa Séptica	19,46	19,93	17,23	19,23	8,78	23,78		21,09	22,50
	Fossa Rudimentar	59,67	41,82	44,03	30,35	46,03	18,55	34,15	45,09	42,95
	Outro Escoadouro	5,19	3,24	3,59	11,17	3,19	8,62	4,77	2,39	1,44
	Sem Instalação Sanitária	7,62	32,62	22,20	29,34	7,83	10,48	2,96	9,93	4,41
II)	Abastecimento de Água em Domicílios (%)	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Rede Geral	54,33	12,19	41,08	48,68	46,64	69,54	59,00	38,01	63,39
	Poço ou Nascente	26,62	43,47	11,86	14,93	40,07	17,05	37,70	52,04	28,76
	Outras	19,05	44,84	47,05	36,39	13,19	13,41	3,30	9,95	7,85
					SÓCIO-ECONÔMICOS					
I)	Analfabeto de 5 anos e mais (%)	19,50	38,40	36,60	26,50	22,20	18,50	17,50	17,40	16,60
II)	PEA em Atividades Industriais (1)	12,40	13,30	16,00	13,00	16,00	16,70	34,20	16,60	23,60
III)	PEA que recebe menos de Cr$ 200,00 (%)	59,40	73,70	64,30	58,70	54,20	40,60	34,80	46,30	46,50

FONTE — IBGE, Censo Demográfico de 1970 — Tabulações Especiais, Rio de Janeiro.
(1) Não inclui construção civil.

de poço também está sujeita à contaminação na medida em que se encontre próximo a instalações sanitárias não ligadas a uma rede geral de esgotos.

Os dados disponíveis quanto às instalações sanitárias mostram que, de modo geral, na maioria das regiões metropolitanas brasileiras é reduzida a população de domicílios que possuem suas instalações sanitárias ligadas à rede geral de esgotos. Predominam as moradias que utilizam a fossa rudimentar, pouco recomendáveis em áreas urbanas onde a densidade demográfica é alta, especialmente onde poços ou nascentes são utilizados como abastecimento de água.

Observa-se que nas regiões metropolitanas do Nordeste, onde a mortalidade é mais elevada, apresentam-se os piores resultados, onde o comum é o predomínio da ausência de instalações sanitárias ou então a presença da fossa rudimentar.

A Região Metropolitana de Fortaleza é entre elas "a que apresenta o quadro mais desfavorável, dos seus 173.339 domicílios, apenas 2,39%, ou seja, 4.148, estão ligados à rede geral de esgoto". Por outro lado, "quando se analisa os dados das regiões metropolitanas do centro-sul brasileiro, que contêm as metrópoles nacionais (São Paulo e Rio de Janeiro) e as metrópoles regionais (Belo Horizonte, Porto Alegre e

Curitiba), observa-se uma sensível melhora em relação às metrópoles antes analisadas, demonstrada pela participação de domicílios ligados à rede geral" [5].

No entanto, embora as regiões metropolitanas do centro-sul estejam em melhor situação, em especial aquelas de maior desenvolvimento econômico e social — Rio de Janeiro e São Paulo — verifica-se, assim mesmo, que a carência de serviços sanitários é ainda um entrave à melhoria das condições de vida de suas populações.

Um outro indicador importante das condições de saneamento dos domicílios refere-se às formas de abastecimento de água. O abastecimento domiciliar através da rede de água encanada tem contribuído substancialmente na redução dos índices de morbidade e de mortalidade por doenças de veiculação hídrica. A esse respeito, afirma-se que "o fato bastante simples, para nos mais afortunados, de setor de água potável, tratada e encanada, pode erradicar, de forma surpreendente, quase todos os tipos de doenças gastrointestinais, quase sempre fatais nos primeiros anos de vida, que engrossam os números da mortalidade infantil (sic)" [6].

A esse respeito, destaque-se, em particular, a situação de Fortaleza, onde a rede geral de água é quase inexistente que, aliada à precariedade das instalações sanitárias, expõe sua população a maior risco de contaminação ambiental, refletindo-se nos níveis mais altos de mortalidade desta região metropolitana.

De um modo geral, constata-se um deficiente nível de atendimento através da rede geral nas regiões metropolitanas brasileiras. As regiões metropolitanas do centro-sul se apresentam também, quanto a este indicador, em melhor situação do que aquelas do Norte e Nordeste, à exceção de Belo Horizonte e Curitiba. Tendo em vista os valores apresentados por estas duas regiões metropolitanas do centro-sul "acredita-se que em Belo Horizonte tais valores sejam explicados pelo desequilíbrio entre oferta deficiente e demanda em crescimento, fruto de seu dinamismo como metrópole de formação recente. Na Região Metropolitana de Curitiba, segundo alguns autores, sua posição como região metropolitana de maior carência do centro-sul, quanto ao serviço ora analisado, decorre da significativa presença de *habitats* rurais, não possuindo portanto este tipo de equipamento muito desenvolvido" [7].

Por último, quanto aos indicadores de situação sócio-econômica (instrução, população economicamente ativa em atividades industriais e população economicamente ativa que recebe menos de Cr$ 200,00) o

[5] NASCIMENTO, M. G. *et alii* — Acessibilidade à Habitação e a Infra-estrutura Domiciliar nas Regiões Metropolitanas Brasileiras. Rio de Janeiro, IBGE/DESPO/DIESM, 1979 (em execução).

[6] OLIVEIRA, Francisco de — A Economia da Saúde. In: O Banquete e o Sonho; Ensaios sobre Economia Brasileira. *Cadernos Debate*, São Paulo, *op. cit.*, nota 31.

[7] NASCIMENTO, M. G. — *op. cit.*

mesmo padrão é encontrado, aparecendo as regiões metropolitanas do centro-sul em melhor situação do que as demais regiões do País.

Em síntese, a análise dos indicadores de nível de vida mostrou que estes estão estreitamente relacionados aos níveis de mortalidade verificados para as regiões metropolitanas.

Por outro lado, embora as regiões metropolitanas mais desenvolvidas do Brasil já tenham alcançado índices de mortalidade mais baixos, é ainda possível uma redução mais substancial desses valores, na medida em que suas populações tenham maiores acessos tanto a serviços de infra-estrutura de saneamento e de abastecimento de água como a melhores níveis de renda.

Nas regiões metropolitanas do Nordeste, no entanto, a situação de suas populações se mostrou pior em todos os níveis, restando tudo a fazer se se quiser diminuir os índices de mortalidade de sua população. Nestas regiões, além da ausência do saneamento, há a péssima remuneração da população com suas conseqüências na capacidade aquisitiva, tendo implicações nas condições alimentares da população onde o problema da desnutrição, "além de poder constituir causa direta de morte, representa fator predisponente e agravante de doenças infecciosas, aumentando substancialmente os coeficientes de morbidade e de fatalidade das mesmas" [8].

3. ESPERANÇA DE VIDA POR RENDA

Neste trabalho se utiliza a renda familiar *per capita* como indicador do nível sócio-econômico dos subgrupos de população. Ainda que este indicador não expresse, em sua totalidade, todo o efeito da classe social sobre a mortalidade, permite, entretanto, evidenciar contrastes associados à condição sócio-econômica do domicílio onde a criança vive. "Antonovsky, em extensa coletânea de estudos internacionais sobre as relações existentes entre as classes sociais, expectativa de vida e mortalidade em geral, conclui que os grupos menos privilegiados têm, de maneira consistente, maiores probabilidades de morrerem mais cedo, se comparados aos grupos mais favorecidos. A despeito da multiplicidade dos métodos e índices utilizados, a evidência estatística, encontrada nos 30 estudos reunidos por Antonovsky, quase que sem exceção dá suporte a esta conclusão. Além disso, também sugere que, à medida que as taxas de mortalidade de uma população caem, diminuem as diferenças de mortalidade entre as classes sociais" [9].

[8] YUNES, J. & RONCHEZES, V. — Evolução de Mortalidade Geral, Infantil e Proporcional no Brasil. In: *Encontro Brasileiro de Estudos Populacionais*, Rio de Janeiro, IBGE, 1976.

[9] ANTONOVSKY, A. — Social class, life expectancy and overall mortality. *The Milbank Memorial Quartely*, 45(1), apr. 1967.

A partir dos dados das tabulações especiais do Censo Demográfico de 1970, foi estimado, para quatro categorias de renda familiar *per capita* e para cada região metropolitana, o número médio de anos de vida ao nascer (l_o). Na tabela 3 estão reunidos os resultados principais observados nas regiões metropolitanas estudadas.

Observa-se um aumento monotônico da esperança de vida com os aumentos da renda. Em relação ao conjunto das regiões metropolitanas como um todo, as famílias que se encontram no estrato de mais alta renda têm uma esperança de vida de 64,85 anos, cerca de doze anos a mais do que os observados para aquelas no estrato de rendimento de Cr$ 1,00 a Cr$ 150,00. Por outro lado, a disparidade de condições de vida entre os grupos de baixa renda e alta renda são bastante marcadas, qualquer que seja a região.

Para qualquer nível da mortalidade, o risco de morrer alcança um máximo nos filhos das mulheres pertencentes ao grupo de renda familiar *per capita* inferior a Cr$ 150,00, decrescendo sistematicamente à medida que aumenta o nível de renda, até alcançar um mínimo nos filhos das mulheres pertencentes ao estrato de renda acima de Cr$ 501,00. As diferenças, no entanto, apesar de serem maiores nas regiões metropolitanas do Nordeste — onde, para a Região Metropolitana de Fortaleza temos uma diferença de 18,70 anos entre os dois extremos, seguindo Salvador (15,30 anos) e Recife (13,65 anos) — são também significativas para as regiões metropolitanas mais desenvolvidas do centro-sul, embora os níveis de esperança de vida alcançados por suas populações sejam mais elevados. Em São Paulo, por exemplo, a diferença entre os dois extremos é de 12,25 anos, embora a esperança de vida alcançada pelos subgrupos de população (66,98 anos no grupo de Cr$ 501,00 e mais e 54,73 anos no de Cr$ 1,00 e Cr$ 150,00) seja bem maior que a verificada para as regiões metropolitanas do Nordeste.

É importante destacar também, aqui, as grandes diferenças encontradas quanto aos níveis de esperança de vida entre os grupos mais pobres (Cr$ 1,00 a Cr$ 150,00) entre as regiões metropolitanas do centro-sul e as do Nordeste. Nestas, as esperanças de vida são bem inferiores. Fortaleza mais uma vez apresenta o menor valor (41 anos), seguida de Recife (46,31 anos) e Salvador (47,55 anos). Nas regiões metropolitanas do centro-sul destaca-se Porto Alegre, onde esse grupo alcança a sua maior esperança de vida (59,39 anos). Nas demais regiões metropolitanas do centro-sul o nível é o mesmo (em torno de 54 anos).

Desta forma, pela análise dos dados, não resta dúvida que a variável *renda familiar per capita* expressa, em boa parte, a posição relativa do grupo familiar em que a criança nasce, em uma escala de bem-estar sócio-econômico. A população pertencente ao estrato de Cr$ 1,00 a Cr$ 150,00, por exemplo, "pertence sem dúvida a um grupo social que tem muito pouco acesso aos bens e serviços, que são o produto social

TABELA 3

VIDA MÉDIA AO NASCER, OBTIDA COM BASE NA MORTALIDADE DOS FILHOS MENORES DE 5 ANOS, RELATIVA ÀS 9 REGIÕES METROPOLITANAS, SEGUNDO A CONDIÇÃO DE NATURALIDADE E A RENDA FAMILIAR *PER CAPITA* — BRASIL — 1970

CLASSES DE RENDIMENTOS FAMILIARES PER CAPITA (Cr$)	Ordem	VIDA MÉDIA AO NASCER				
		Conjunto Metropolitano do Nordeste	Regiões Metropolitanas			
			Belém	Fortaleza	Recife	Salvador
TOTAL						
Total		54,03	55,66	41,81	47,05	48,20
Até 150	1	52,46	54,56	41,00	46,31	47,55
151—300	2	60,31	61,46	57,71	55,64	51,63
301—500	3	63,31	62,20	59,70	59,73	56,17
501 e mais	4	64,85	—	—	59,96	62,85
	(4—1)	12,39	7,64	18,70	13,65	15,30
NATURAIS						
Total		55,33	55,74	43,09	48,12	48,84
Até 150	1	53,33	55,79	42,25	47,45	48,35
151—300	2	62,51	61,14	60,48	57,38	51,44
301—500	3	65,30	62,07	60,50	61,71	56,44
501 e mais	4	66,49	—	—	62,45	65,10
	(4—1)	13.16	6,28	18,25	15,00	16,75
MIGRANTES						
Total		53,09	53,69	40,26	45,88	47,37
Até 150	1	51,51	52,85	39,33	45,08	46,62
151—300	2	58,46	62,39	54,97	54,39	51,51
300—500	3	61,22	62,94	59,24	58,12	55,45
501 e mais	4	62,92	—	—	58,22	61,71
	(4—1)	11,41	10,09	19,91	13,14	15,09

CLASSES DE RENDIMENTOS FAMILIARES PER CAPITA (Cr$)	Ordem	VIDA MÉDIA AO NASCER					
		Conjunto Metropolitano do Centro-sul	Regiões Metropolitanas				
			Belo Horizonte	Rio de Janeiro	São Paulo	Curitiba	Porto Alegre
TOTAL							
Total		54,03	52,99	56,00	56,58	55,66	60,51
Até 150	1	52,46	52,04	54,71	54,73	54,39	59,39
151—300	2	60,31	60,50	60,07	60,90	61,49	65,54
301—500	3	63,31	61,02	63,66	64,27	64,07	66,54
501 e mais	4	64,85	—	65,32	66,98	65,51	67,53
	(4—1)	12,39	8,98	10,61	12,25	11,12	8,14
NATURAIS							
Total		55,33	56,53	59,17	61,31	56,31	61,41
Até 150	1	53,33	55,47	57,69	58,26	55,38	60,29
151—300	2	62,51	64,65	62,51	64,22	61,09	65,32
301—500	3	65,30	—	65,96	66,86	63,63	66,57
501 e mais	4	66,49	—	68,14	67,80	65,22	68,85
	(4—1)	13,16	—	10,45	9,54	9,84	8,56
MIGRANTES							
Total		53,09	51,68	54,05	55,54	54,99	59,95
Até 150	1	51,51	50,93	53,14	53,74	53,55	58,90
151—300	2	58,46	58,21	57,57	58,78	61,68	65,74
301—500	3	61,22	59,70	60,83	61,87	64,46	66,74
501 e mais	4	62,92	—	62,27	65,83	66,34	67,47
	(4—1)	11,41	8,77	9,13	12,09	12,79	8,57

FONTE — Estimativas utilizando o método de Brass, obtidas tomando por base: IBGE, Censo Demográfico de 1970. Tabulações Especiais, Rio de Janeiro.

do trabalho do homem. Vivem, por isso, em um ambiente físico, biológico e social em extremo hostil ao desenvolvimento normal da criança e a sua própria sobrevivência" [10]. Só a título de ilustração, as crianças nordestinas nascidas neste grupo de renda familiar vivem em condições de mortalidade só comparáveis aos existentes na Europa ocidental em 1860. Para aquilatar a significação deste fato, reportemo-nos de novo a alguns indicadores de saneamento, só que agora para cada classe de rendimento.

Nas tabelas 4 e 5 tem-se, respectivamente, para cada região metropolitana a percentagem de domicílios quanto à forma de abastecimento de água utilizada e quanto aos tipos de instalações sanitárias, segundo classes de rendimentos.

De modo geral, os dados disponíveis são bastante categóricos. A possibilidade de eliminação dos dejetos e detritos, que é essencial para a melhoria do estado sanitário das populações, está ausente na maioria dos domicílios com rendimentos inferiores a 5 salários mínimos nas regiões metropolitanas do Nordeste. Mesmo na classe de rendimentos superiores a 5 salários mínimos, é baixa a proporção dos domicílios cujas instalações sanitárias se encontram ligadas à rede geral. Dentre as regiões metropolitanas, Recife aparenta melhor situação quanto a este indicador, principalmente na última classe de renda. Desta forma, à medida em que diminui o rendimento domiciliar há um predomínio da fossa rudimentar e ausência de instalações nos domicílios, o que, de certa forma, explica as altas taxas de mortalidade encontradas para os estratos de rendimentos inferiores, devido principalmente às doenças infecciosas e transmissíveis, dado a carência de saneamento ambiental. Da mesma forma, quando se considera o abastecimento de água em domicílios através da rede geral, verifica-se também um deficiente nível de atendimento, principalmente para os estratos inferiores de renda. A Região Metropolitana de Fortaleza, dentre as regiões metropolitanas do Nordeste, apresenta os mesmos níveis de abastecimento, mesmo no estrato mais alto de renda (apenas 50,16% dos domicílios ligados à rede geral).

Por outro lado, quando se analisa as mesmas variáveis para as regiões metropolitanas mais desenvolvidas, a tendência por renda é a mesma verificada para o Nordeste, embora se apresentem melhor servidas quanto a esses tipos de serviços.

De qualquer forma, são as populações mais carentes que se encontram em pior situação. Nas Regiões Metropolitanas do Rio de Janeiro e São Paulo, de maior adensamento populacional, há um desequilíbrio entre a oferta e demanda de serviços de infra-estrutura de saneamento

[10] BEHM, Hugo & PRIMANTE, D. A. — Mortalidad en los primeros años de vida en la America Latina. *NOTAS DE POBLACIÓN*. Revista Latino Americana de Demografia. San José (Costa Rica), 6(16): 23-44, abr. 1978.

TABELA 4

PERCENTAGEM DE DOMICÍLIOS QUANTO À FORMA DE ABASTECIMENTO DE ÁGUA UTILIZADA, SEGUNDO CLASSES DE RENDIMENTOS

REGIÕES METROPOLITANAS — BRASIL — 1970

FORMAS DE ABASTECIMENTO E CLASSES DE RENDA (Salários mínimos)	PERCENTAGEM DOS DOMICÍLIOS NAS REGIÕES METROPOLITANAS DE								
	Belém	Fortaleza	Recife	Salvador	Belo Horizonte	Rio de Janeiro	São Paulo	Curitiba	Porto Alegre
ATÉ 1 S.M.									
Rede geral	31,52	2,49	16,73	23,21	22,00	39,67	28,78	11,09	32,60
Poço ou nascente	36,09	33,67	12,74	19,66	50,37	31,46	60,09	58,31	47,18
Outros	32,39	63,84	70,53	57,13	27,63	28,87	11,13	30,60	20,22
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
DE 1 A 3 S.M.									
Rede geral	49,28	9,25	40,28	42,02	36,32	59,45	41,87	24,78	53,08
Poço ou nascente	30,12	50,63	13,63	16,80	49,55	22,87	53,15	65,54	37,58
Outros	20,60	40,12	46,08	41,18	14,13	17,68	4,98	9,68	9,34
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
DE 3 A 5 S.M.									
Rede geral	71,52	24,02	73,45	67,66	60,09	79,55	61,78	47,16	75,63
Poço ou nascente	19,01	57,49	8,90	11,58	34,94	12,29	36,34	50,24	21,13
Outros	9,47	18,49	17,65	20,76	4,97	8,16	1,88	2,60	3,24
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5 S.M. E MAIS									
Rede geral	90,43	50,16	93,62	89,91	86,05	94,35	84,65	77,48	92,74
Poço ou nascente	7,14	45,27	3,74	4,92	12,80	3,78	14,85	22,03	6,62
Outros	2,43	4,57	2,64	5,17	1,15	1,87	0,50	0,49	0,64
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

FONTE — IBGE, Censo Demográfico 1970, Tabulações Especiais, Rio de Janeiro (dados tirados do trabalho de NASCIMENTO, Ma.G.; FREITAS, A.L.; MONTEIRO, V.S. e BARBOSA, J.G. — Acessibilidade à Habitação e à Infra-Estrutura Domiciliar nas Regiões Metropolitanas Brasileiras — IBGE/DESPO/DIESM, 1979 (trabalho em andamento).

básico, motivado, em grande parte, pelo seu crescimento desordenado. "O Plano Nacional de Saneamento, elaborado pelo BNH para tratar da questão do saneamento e da rede de água, vem-se mostrando ineficiente para enfrentar tal situação. Mais uma vez os preços cobrados por tais serviços deixam de fora as populações mais pobres, que não podem arcar com as despesas da ligação de suas residências à rede geral. E ao invés de elaborar planos mais simplificados de eliminação de dejetos, o *PLANASA* atém-se às normas e padrões técnicos a que obedecem os países mais desenvolvidos, encarecendo a obra e atendendo, conseqüentemente, apenas às parcelas mais bem providas da população" [11].

Assim, por exemplo, no caso do abastecimento de água, em algumas áreas "a instalação de chafarizes públicos como etapa inicial na introdução gradativa do abastecimento de água parece ser uma medida mais

[11] PAULA, S. G. — Saúde em Áreas Urbanas. In: *Revista de Administração Pública*. Rio de Janeiro, 12(2): 163-82, abr./jun., 1978.

TABELA 5

PERCENTAGEM DE DOMICÍLIOS QUANTO AOS TIPOS DE INSTALAÇÕES SANITÁRIAS, SEGUNDO CLASSES DE RENDIMENTOS

REGIÕES METROPOLITANAS — BRASIL — 1970

CLASSE DE RENDIMENTOS POR TIPOS DE INSTALAÇÕES SANITÁRIAS (Salários mínimos)	PERCENTAGEM DOS DOMICÍLIOS NAS REGIÕES METROPOLITANAS DE								
	Belém	Fortaleza	Recife	Salvador	Belo Horizonte	Rio de Janeiro	São Paulo	Curitiba	Porto Alegre
ATÉ 1 S.M.									
Rede geral	2,36	0,30	2,26	2,95	9,44	12,40	}	4,31	6,82
Fossa séptica	5,75	4,76	5,82	5,55	3,94	17,26	} 25,87	6,41	8,85
Fossa rudimentar	69,81	35,22	48,55	28,60	64,56	29,97	54,42	52,20	70,57
Outros	7,78	4,84	4,17	10,49	3,39	13,41	6,77	3,72	2,04
Não tem	14,30	54,88	39,20	52,39	18,65	26,96	12,94	33,36	11,72
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
DE 1 A 3 S.M.									
Rede geral	4,07	1,27	8,49	6,02	21,62	22,92	}	11,06	16,10
Fossa séptica	13,86	17,98	17,89	13,49	9,42	27,08	} 41,39	17,85	19,36
Fossa rudimentar	68,68	54,64	52,47	36,94	57,49	25,22	48,20	59,56	57,64
Outros	5,67	2,79	3,86	11,88	3,82	11,58	6,11	2,80	1,81
Não tem	7,71	23,32	17,29	31,64	7,65	13,20	4,30	8,73	5,09
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

CLASSE DE RENDIMENTOS POR TIPOS DE INSTALAÇÕES SANITÁRIAS (Salários mínimos)	PERCENTAGEM DOS DOMICÍLIOS NAS REGIÕES METROPOLITANAS DE								
	Belém	Fortaleza	Recife	Salvador	Belo Horizonte	Rio de Janeiro	São Paulo	Curitiba	Porto Alegre
DE 3 A 5 S.M.									
Rede geral	9,12	5,09	22,91	14,98	47,04	42,81	}	24,89	34,13
Fossa séptica	33,19	43,32	36,15	27,81	14,54	30,32	} 61,83	31,34	32,42
Fossa rudimentar	50,85	44,85	34,18	32,77	32,91	14,59	31,99	39,52	30,39
Outros	3,58	1,16	2,51	12,95	3,47	7,19	4,92	1,99	1,15
Não tem	3,26	5,58	4,25	11,47	2,01	5,09	1,26	2,26	1,91
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5 S.M. E MAIS									
Rede geral	27,67	12,03	51,42	25,43	78,40	73,83		52,12	61,53
Fossa séptica	46,43	67,13	35,44	47,31	9,01	17,93	83,12	31,65	29,79
Fossa rudimentar	23,86	19,46	11,42	15,93	10,74	4,64	13,98	14,81	7,74
Outros	1,31	0,36	0,91	9,04	1,44	2,53	2,59	0,99	0,56
Não tem	0,73	1,02	0,81	2,26	9,41	1,07	0,31	0,43	0,38
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

FONTE — IBGE, Censo Demográfico 1970, Tabulações Especiais, Rio de Janeiro (dados tirados do trabalho de NASCIMENTO, Ma.G.; FREITAS, A.L.; MONTEIRO, V.S. e BARBOSA, J.G. — "Acessibilidade à Habitação e à Infra-Estrutura Domiciliar nas Regiões Metropolitanas Brasileiras" — IBGE/DESPO DIESM, 1979 (trabalho em andamento).

viável do que a instalação de uma rede geral em ligações individuais, uma vez que é opção mais acessível à população de baixa renda" [12].

Entretanto, a necessidade de tornar rentáveis os projetos e dado que o assentamento das camadas mais pobres é nitidamente ineficiente, "a ação governamental quase sempre limita-se a seguir a dinâmica implantada pelo setor privado, estando os investimentos públicos a serviço

[12] NASCIMENTO, M. G. — *op. cit.* p. 34.

da valorização e da especulação realizadas pelo setor imobiliário" [13]. E mesmo quando faz investimentos nas áreas onde mora a população de baixa renda, o aumento dos aluguéis, impostos locais ou valor do solo podem resultar na expulsão da população de baixa renda [14].

Em resumo, vimos como a ausência de saneamento básico está afetando os níveis de vida das populações de baixa renda, em especial no Nordeste, onde este serviço é quase inexistente para estas populações. Por outro lado, fica evidenciado que a oferta destes serviços não está voltada para a melhoria das condições de vida das populações mais necessitadas; este fato aliado às políticas concentradoras de renda e de benefícios está influindo decisivamente sobre o padrão de saúde das populações de baixa renda, que se reflete nas altas taxas de mortalidade encontradas para as populações mais pobres. Desta forma, os problemas de saúde, sobretudo os de saúde pública nas regiões menos desenvolvidas, são decorrentes dos baixos níveis de rendimento e das miseráveis condições de vida de suas populações. As doenças de massa decorrentes das péssimas condições de moradia e saneamento só serão combatidas de maneira mais eficaz com melhor alimentação e melhor distribuição da renda, assegurando, assim, melhores condições econômicas e emprego.

Neste sentido, ilustramos a seguir, com base nos dados do Estudo Nacional de Despesa Familiar (ENDEF), as condições de nutrição das populações nas Regiões Metropolitanas de Porto Alegre (de baixa mortalidade) e de Recife (de alta mortalidade), vendo como as diferenças de poder aquisitivo familiar atuam sobre o consumo calórico de suas populações. Para tanto, baseamo-nos no trabalho de Lustosa [15].

Trabalhando com a variável *despesa corrente anual per capita* como *proxy* da renda, estratifica-se as famílias computando os quartis de despesa corrente *per capita* em cada região metropolitana, "com o propósito de contrastar o perfil de consumo alimentar de subconjuntos das populações que guardam entre si a mesma relação quanto à distribuição de despesa corrente *per capita*" [16].

Nas tabelas 6 e 7 apresenta-se o consumo calórico por comensal-adulto-dia, por clases de despesa corrente *per capita* para as duas regiões metropolitanas.

Evidencia-se, inicialmente, a diferença entre os níveis de renda das populações de Porto Alegre e Recife, que se traduz nos níveis dos quartis de despesa corrente *per capita* observados. Em Porto Alegre, por exemplo, os valores dos limites superiores das classes de despesa correspondem

[13] PAULA, S. G. — *op. cit.* p. 180.

[14] VETTER, David Michael *et alli* — Espaço, valor da terra e eqüidade dos investimentos em infra-estrutura urbana: uma análise do município do Rio de Janeiro. Rio de Janeiro, IBGE/DEISO/, 1979 (mimeo, ver 4-12).

[15] LUSTOSA, T. Q. O. — Perfil nutricional das populações de Porto Alegre e Recife. Rio de Janeiro, IBGE/DESCO, 1979 (mimeo-inédito).

[16] Ibid. *op. cit.* p. 33.

TABELA 6

CONSUMO CALÓRICO POR COMENSAL-ADULTO-DIA, POR CLASSES DE DESPESA CORRENTE PER CAPITA FAMILIAR *, SEGUNDO GRUPOS DE ALIMENTOS REGIÃO METROPOLITANA DE PORTO ALEGRE — 1974/75

ALIMENTOS (K/Cal)	CONSUMO CALÓRICO POR COMENSAL-ADULTO-DIA				
	Todas as Classes	Classes de despesa corrente per capita (Cr$)			
		Menos de 4 047	De 4 047 a 6 465	De 6 466 a 11 766	Acima de 11 766
Cereais e derivados	1 173	1 236	1 217	1 144	1 017
Óleos e gorduras	431	398	455	458	440
Açúcar e derivados	378	405	396	356	326
Carnes e pescados	327	207	328	412	459
Ovos, leite e queijos	262	158	246	328	411
Leguminosas e oleaginosas	172	210	176	152	114
Tubérculos, raízes e similares	103	99	111	106	93
Frutas	67	41	68	78	104
Bebidas e derivados	51	23	37	67	107
Legumes e verduras	36	24	36	42	52
TOTAL	3 000	2 801	3 060	3 143	3 123

FONTE — IBGE-ENDEF, 1974/75, Tabulações Especiais, Rio de Janeiro (tabela extraída do trabalho de LUSTOSA, T.Q.O.-op. citado, p.40).

*É a seguinte a correspondência em termos de salário mínimo anual (SMA).

1) menos de 4 047 — menos de 0,9 SMA.
2) de 4 047 a 6 465 — de 0,9 SMA a 1,5 SMA.
3) de 6 465 a 11 766 — de 1,5 SMA a 2,6 SMA.
4) acima de 11 766 — acima de 2,6 SMA.

a aproximadamente o dobro do encontrado para Recife. Constata-se, de imediato, o maior consumo de calorias em Porto Alegre em todos os estratos das populações, sendo que a magnitude dessa diferença em relação a Recife se reduz do primeiro estrato (40%) para o quarto estrato. Observa-se também que as maiores disparidades no consumo de calorias entre os estratos ocorrem em Recife. Nesta região a classe de maior poder aquisitivo consome mais 41% do que o estrato mais pobre, enquanto em Porto Alegre a discrepância é de apenas 12%.

Estes resultados refletem, assim, as diferenças nos níveis de mortalidade entre os de renda mais baixa e mais alta. Como vimos, a diferença era de 13,65 anos em Recife e de apenas 8,14 anos em Porto Alegre (tabela 3).

Ainda do mesmo estudo, a autora conclui que a dieta média da população de Recife apresenta qualidade nutricional inferior à da população de Porto Alegre. "Os requerimentos calóricos médios não são cobertos para o estrato de menor poder aquisitivo e 48% das famílias

TABELA 7

CONSUMO CALÓRICO POR COMENSAL-ADULTO-DIA, POR CLASSES DE DESPESA CORRENTE PER CAPITA FAMILIAR *, SEGUNDO GRUPOS DE ALIMENTOS REGIÃO METROPOLITANA DE RECIFE — 1974/75

ALIMENTOS (K/Cal)	CONSUMO CALÓRICO POR COMENSAL-ADULTO-DIA				
	Todas as Classes	Classes de despesa corrente per capita (Cr$)			
		Menos de 1 779	De 1 179 a 3 114	De 3 115 a 6 135	Acima de 6 135
Cereais e derivados	870	710	884	966	965
Açúcar e derivados	387	309	388	447	424
Tubérculos, raízes e similares	333	452	363	284	196
Carnes e pescados	258	163	231	300	370
Leguminosas e oleaginosas	226	211	234	252	208
Óleos e gorduras	169	72	147	213	273
Ovos, leite e queijos	134	69	93	148	250
Frutas	75	36	60	94	121
Bebidas e derivados	29	14	22	34	48
Legumes e verduras	24	11	20	30	39
TOTAL	2 504	2 047	2 442	2 768	2 894

FONTE — IBGE-ENDEF, 1974/75, Tabulações Especiais, Rio de Janeiro (tabela extraída do trabalho de LUSTOSA, T.Q.O. — op. citado, p.41).

*É a seguinte a correspondência em termos de Salário Mínimo Anual (SMA).

1) menos de 1 779 — menos de 0,4 SMA.
2) de 1 779 a 3 114 — de 0,4 SMA a 0,7 SMA.
3) de 3 115 a 6 135 — de 0,7 SMA a 1,4 SMA.
4) acima de 6 135 — acima de 1,4 SMA.

compreendidas nesse estrato não alcança uma taxa de adequação de 100%. Quanto à vitamina A, observa-se que no primeiro estrato da população de Recife a taxa de adequação média é inferior a 100% e 71% das famílias aí alocadas não cobrem seus requerimentos. Para as populações como um todo, 42% das famílias observadas em Recife apresentam taxas de adequação inferior a 100%, sendo essa proporção, em Porto Alegre, 30%" [17].

Assim, a melhor situação nutricional identificada, em geral, para os estratos de população de Porto Alegre pode, portanto, ser imputada a maiores níveis de renda e à sua distribuição mais eqüitativa nessa área, refletindo-se nos níveis mais baixos de mortalidade nesta região.

Do exposto, conclui-se que para melhorar a qualidade de vida das populações de baixa renda (principalmente no Nordeste), não basta apoiar-se somente em medidas médico-sanitárias, uma vez que a maioria das endemias que atingem esse estrato estão estreitamente correlacio-

[17] Ibid, *op. cit.* p. 69.

nadas ao estado de miséria em que vivem essas populações. Os serviços de saneamento básico (água e esgoto), embora necessários, de um modo geral, não terão resultados duradouros se não forem tomadas medidas paralelas que modifiquem a estrutura econômico-social e, dessa maneira, as condições alimentares da população.

Finalmente, neste tópico do trabalho, procede-se agora de forma sumária à comparação entre os níveis de mortalidade por renda para os filhos de migrantes e naturais, tomando como base os dados da tabela 3.

Em linhas gerais, a mortalidade dos migrantes, como dos naturais, segue o mesmo padrão já delineado, ou seja, mortalidade mais alta nos estratos de renda inferior, descendo à medida que aumenta a renda familiar *per capita*. Nas regiões metropolitanas do centro-sul é onde se verificam as mais baixas taxas de mortalidade, conforme pode ser visto na tabela, bem como as menores diferenças entre as mortalidades nos estratos baixo e alto de renda.

Por outro lado, quando se comparam os níveis de mortalidade para as duas populações, observa-se, de um modo geral, que os filhos de naturais sobrevivem mais que os filhos de migrantes em todos os estratos de renda e em todas as regiões metropolitanas. No entanto, as diferenças só são significativas apenas para as Regiões Metropolitanas do Rio de Janeiro, São Paulo e Belo Horizonte. Nas demais são irrelevantes as diferenças. Um outro fato importante a ser retirado das tabelas consiste na comparação entre os níveis mais baixos de mortalidade dos migrantes residentes nas regiões metropolitanas do centro-sul com as populações (migrante e natural) residentes nas regiões metropolitanas do Nordeste. Como se sabe, é grande o peso de migrantes nordestinos, principalmente nas metrópoles nacionais — Rio de Janeiro e São Paulo — o que nos leva a concluir que esta população encontraria nestas regiões metropolitanas melhores condições de sobrevivência, refletindo-se nos níveis de mortalidade, relativamente mais baixos que os das populações das áreas de origem.

Por outro lado, as maiores diferenças que são encontradas entre os níveis de mortalidade nas metrópoles nacionais para as duas populações, de certa maneira, ainda refletem parte da alta mortalidade dos migrantes nordestinos que, ao migrarem, levariam o padrão de mortalidade prevalecente na área de origem. No entanto, como a fixação nas áreas de destino é bastante seletiva no sentido de que são "sobreviventes" de um processo que já durou certo período de tempo, onde a condição de não adaptação, para muitos, determinou a reemigração para outras áreas ou até o retorno [18], vão com o tempo adquirindo o padrão existente na área de residência.

[18] MARTINE, G. — Adaptação dos Migrantes ou Sobrevivência dos mais Fortes. *Relatório Técnico, 30, Projeto de Planejamento de Recursos Humanos.* Brasília (DF), 1976.

UMA BREVE INTRODUÇÃO À ANÁLISE ESTATÍSTICA COM SPSS

(Statistical Package for the Social Sciences) *

David Michael Vetter **

SUMÁRIO

1. INTRODUÇÃO

Segundo Claudio de Moura Castro (1977, p. 94), "os cursos usuais de Estatística deixam os pesquisadores insuficientemente equipados para

* O Autor agradece as sugestões de Nicia M. Bessa, Maurício F. Mendonça de Aguiar, Luís Otávio F. Barreto Leite, Rosa Maria Ramalho Massena, Luís Carlos V. dos Santos, Ricardo Luís Cardoso e José Roberto de Sousa Santos, e, ainda, a seus alunos do Mestrado em Educação da PUC/Rio de Janeiro. Independentemente desta valiosa colaboração, as idéias desenvolvidas neste trabalho são de inteira responsabilidade do autor.

** Departamento de Estudos e Indicadores Sociais — DEISO/SUEGE/IBGE. Professor do Mestrado de Educação, PUC/Rio.

enfrentar os problemas e contradições nas investigações em ciências sociais". Uma parte do problema é que "a estatística é reconstruída a partir de uma seqüência lógico-dedutiva", que geralmente não inclui a prática com o teste de hipóteses científicas de uma pesquisa social. Com efeito, este último é, via de regra, apresentado nos cursos de metodologia, mas mesmo neles sem a necessária prática da análise de dados concretos de uma pesquisa real. Por estas razões, dentre os alunos dos cursos de ciências sociais que estudam Estatística, são poucos os que aprendem a aplicá-la na prática e, assim sendo, acabam por esquecer tudo rapidamente. Em geral, é apenas na hora de começar a realizar sua tese que o aluno se dá conta da falta deste conhecimento prático.

Na minha opinião, os programas "enlatados" de análise estatística poderiam ser de grande utilidade na resolução deste problema, uma vez que eles permitiriam ao aluno testar hipóteses elaboradas por ele com dados de uma pesquisa em sua área de interesse. Além de mostrar como aplicar os diferentes métodos, esta prática normalmente estimula o interesse dos alunos, já que comprova sua utilidade na pesquisa social e reduz o "medo" que muitos manifestam em relação ao computador e à utilização de métodos quantitativos. Assim sendo, um curso de métodos quantitativos com um sistema de programas estatísticos poderia servir como uma "ponte" entre os cursos de estatística e de metodologia de pesquisa, contribuindo, desse modo, para a resolução deste sério problema identificado por Castro (ibidem).

O SPSS (*Statistical Package for the Social Sciences*) oferece uma série de vantagens para este tipo de curso, uma vez que ele foi concebido especialmente para satisfazer aos requisitos da análise estatística aplicada às ciências sociais. E ainda é extremamente fácil de ser utilizado, inclusive pelo usuário que não tenha conhecimento de computação. Além disso, possui grande flexibilidade, o que permite a modificação e criação de variáveis, manipulação de arquivos e uma grande diversidade de tipos de análise estatística, incluindo:

1. estatísticas descritivas para a população e subpopulação;
2. distribuição de freqüência e histogramas;
3. tabulações cruzadas;
4. correlação de Pearson, a não-paramétrica e a parcial;
5. análise de variância e de covariância;
6. análise fatorial;
7. análise discriminante;
8. regressão simples e múltipla, e
9. correlação canônica.

Devido à flexibilidade do SPSS, bem como à facilidade de seu uso e à sua abrangência, há mais de 1.000 instalações no mundo que se servem dele. No Brasil muitas instituições utilizam o sistema, tais como as Universidades Federais do Rio de Janeiro e do Rio Grande do Sul, a PUC-Rio, o Control Data Corporation, o SERPRO e o IBGE.

O presente trabalho tem, então, por finalidade principal mostrar como fazer análise estatística com SPSS. Espero que, dessa forma, ele enseje um aumento na análise de dados de pesquisas reais por parte de alunos e outros. Considerando-se que o objetivo é primordialmente didático, todo esforço será feito para que a exposição seja a mais simples e concisa possível. Serão tratados aqui somente os métodos para uma ou duas variáveis, deixando-se os métodos multivariados para um trabalho posterior.

Acreditamos que a melhor maneira de se aprender a utilizar SPSS (ou qualquer outra linguagem de computador) seja através da prática, começando-se por problemas simples e passando-se depois aos mais complexos. Temos notado que muitos alunos encontram grandes dificuldades na etapa inicial da aprendizagem, perdendo-se nas 675 páginas do Manual do SPSS e, por isso, não conseguindo preparar seu primeiro programa, ou conduzindo-o com muita dificuldade. Visando à solução deste problema, procuraremos mostrar apenas o essencial para se definir, modificar e criar variáveis com SPSS e utilizar seus procedimentos estatísticos mais comumente empregados, ilustrando ainda como se proceder a cada operação, através de um exemplo simples.

Para descrições mais detalhadas, o aluno deve consultar o Manual do SPSS:

N. H. Nie *et alii,* SPSS: *STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES (NEW YORK: McGraw-Hill, 1975).*

Tentaremos sempre indicar as páginas onde o aluno pode encontrar estas informações, utilizando *SPSS* como a abreviatura do título deste Manual. Existe ainda um manual introdutório, que é mais resumido:

W. R. Klecka *et alii* SPSS PRIMER (*New York McGraw-Hill, 1975*).

A diferença entre o tratamento aqui escolhido e o encontrado em textos sobre SPSS em português é que tentaremos ir além das instruções de como se elaborar o programa de SPSS para dar uma idéia muito geral de como interpretar os resultados gerados por ele, inclusive citando outras fontes onde o aluno pode achar maiores informações sobre as técnicas utilizadas (ver a bibliografia). Para os que querem somente uma rápida introdução aos cartões de controle de SPSS, recomendamos o texto encontrado no sistema de documentação do Rio Data Centro (LISTADOC) da PUC/Rio.

A fim de exemplificarmos as diversas operações do SPSS, analisaremos os níveis de rendimento dos alunos do Mestrado em Educação nos exames de seleção em Inglês, Estatística e Redação nas turmas de 1976, 1977 e 1978. Foi selecionada uma amostra aleatória simples de 90 casos, sendo 30 de cada área de especialização. As variáveis escolhidas foram: data de nascimento, sexo, área de especialização, turma e as notas das três provas. Analisaremos as relações entre as notas obtidas nessas provas e as outras variáveis (ver 7.0 para uma listagem do deck).

2. ALGUNS CONCEITOS BÁSICOS DE ANÁLISE ESTATÍSTICA

Antes de procedermos à discussão do SPSS, discutiremos alguns conceitos básicos necessários para o entendimento de como se realiza a análise estatística com SPSS.

2.1 Variáveis Contínuas e Categoriais (Não-Contínuas)

Rodrigues (1975, p. 35) adota a definição de variável "como sendo uma propriedade à qual se atribuem valores numéricos, valores estes suscetíveis de variação ao longo de uma amplitude finita (variável categorial [1] ou não-contínua) ou infinita (variável contínua)". Como exemplos de variáveis comuns em psicologia e educação, ele enumera um certo número de propriedades que variam de indivíduo para indivíduo e circunstância para circunstância, tais como: inteligência, sexo, preconceito, destreza manual, habilidade numérica, conformismo, atividade, agressividade, rendimento escolar, sociabilidade, introversão e um sem-número de outras. É de interesse científico saber se ao modificar-se uma variável (digamos, grau de inteligência), uma outra também se modifica (digamos, rendimento escolar)" (p. 35). São utilizados métodos de investigação científica para analisar a natureza das relações entre as variáveis escolhidas como relevantes.

Como foi exposto acima, uma variável contínua varia ao longo de uma amplitude infinita, enquanto uma variável categorial (não-contínua) tem um número finito de categorias. Exemplos de variáveis contínuas incluem idade, peso, escores de um teste de inteligência, etc. A variável idade é contínua, sendo que, teoricamente, ela pode assumir qualquer valor entre dois dados. A idade de um indivíduo pode ser 23 anos, 23 anos, 3 meses e 15 dias; 23 anos, 6 meses e 20 dias, etc. A altura de um indivíduo é também uma variável contínua, podendo corresponder a 1,90 metros, 1,94 metros, 1,943 metros, e assim por diante.

Existem dois tipos de variáveis categoriais ou não-contínuas. O primeiro pode ser exemplificado por um número finito de categorias

[1] Outros autores usam, ao invés de *categorial*, o termo *categorizada*.

nominais que são mutuamente exclusivas, tais como: sexo, estado civil, raça, cidade natal, etc. O aluno pode ser, por exemplo, do sexo (1) masculino ou (2) feminino, e do estado civil (1) solteiro, (2) casado, etc. Note-se que os números são utilizados nestes casos apenas para identificação das categorias e não para mostrar a distância entre estas ou a ordem das mesmas. Quanto ao segundo tipo, é possível considerar-se que uma variável contínua é transformada em uma variável não-contínua, através da agregação ou agrupamento dos valores em classes ou grupos (ver Spiegel, p. 43-69). Poderíamos, por exemplo, agrupar os alunos segundo grupos de idade:

(1) até 24 anos
(2) de 25 a 29 anos
(3) de 30 a 39 anos
(4) de 40 ou mais anos.

No SPSS estes números de identificação das categorias de uma variável não-contínua (categorial) são chamados *values* (valores).

Às vezes a classificação das variáveis é difícil, como no nosso caso dos escores das provas de redação, estatística e inglês, onde os valores podem teoricamente situar-se entre zero e dez (4,0, 4,5, 4,8, etc.), mas foram arredondados para o número inteiro mais próximo. Neste caso, temos uma variável teoricamente contínua, mas com um máximo de 11 grupos (de 0 a 10). Por isso, poderíamos tratar estas três variáveis como se fossem agrupadas, ou como variáveis contínuas.

2.2 Níveis de Medida

A classificação de variáveis segundo seu nível de medida é extremamente importante porque determina os tipos de métodos estatísticos que podem ser utilizados. O nível de medida de uma variável depende de sua capacidade de mensurar a distância entre os valores de uma variável ou a ordem deles (*SPSS*, p. 4-6).

2.2.1 O Nível de Medida Nominal

Com o nível de medida nominal, "números são atribuídos a categorias mutuamente exclusivas" (Rodrigues, p. 21). Mas estes números são utilizados apenas para identificação das categorias (como rótulos ou símbolos) e não dão nenhuma informação sobre a ordenação ou a distância entre estas categorias. Seria absurdo, por exemplo, pensar na ordenação de alunos segundo seu estado civil ou sexo.

2.2.2 *O Nível de Medida Ordinal*

A escala ordinal mostra a ordenação das categorias, mas não a distância entre elas. Por exemplo, poderíamos classificar alunos em grupos segundo suas notas: alto, médio e baixo. "Nota-se quanto ao caso de consideração dos grupos em termos de escala ordinal que se podem fazer comparações em termos de maior e menor, ou superior e inferior; não se pode, porém, estabelecer relações de quantas vezes maior ou menor ou de quantas vezes superior ou inferior. Seria absurdo dizer, por exemplo, que um estudante situado no percentil 90 é três vezes mais inteligente do que um situado no percentil 30, ou que um estudante que tenha atingido o percentil 70 sabe duas vezes mais que um reprovado, que se situou no percentil 35" (Rodrigues, p. 23).

2.2.3 *Os Níveis de Medida Intervalar e de Razão*

Com a escala intervalar sabemos, além da ordenação dos valores, as distâncias entre eles em termos de unidades fixas e iguais (tais como graus de temperatura ou valores em cruzeiros). Os escores padronizados de um teste é um exemplo de uma escala ou nível de medida intervalar. Neste caso, poderíamos dizer que a diferença entre os escores de 60 e 70 é igual à diferença entre os escores de 30 e 40 (Rodrigues, p. 23-24).

Existe um debate de longa duração nas ciências sociais sobre o uso dos métodos paramétricos na análise dos escores de testes de inteligência, aptidão, etc. Embora estes escores sejam do nível de mensuração ordinal *stricto sensu,* muitos pesquisadores defendem a posição de que eles podem ser tratados como se fossem de nível intervalar, sem maiores distorções e, então, ser analisados com métodos paramétricos. A coletânea de Kirk (1972, p. 47-79) apresenta trabalhos que ilustram os dois lados desta questão. Aqui as notas serão analisadas como se fossem de nível de mensuração intervalar.

Embora a escala de medida intervalar permita comparações das *diferenças* entre valores, ela *não* permite comparações entre as magnitudes destes valores, dado que não existe um zero absoluto, ou seja, um valor que possa indicar a ausência de distância. A escala de razão depende da existência de um zero absoluto que representa a ausência de distância entre os valores de uma variável, como no caso de medidas de distância física. Em virtude da falta de variáveis que têm um zero absoluto, nas pesquisas estatísticas de interesse para as ciências sociais, a escala de razão é muito pouco utilizada.

2.2.4 *Um Caso Especial: A Variável Binária ou Dicotomizada*

Por causa de suas propriedades matemáticas, uma variável binária pode ser tratada como se fosse uma medida com um nível intervalar.

O sexo, por exemplo, poderia ser definido na forma de uma variável binária:

Sexo = 1, se for masculino
= 0, se não for masculino.

Poderíamos ordenar as variáveis (não importando qual seja a maior e qual seja a menor) e saberíamos a distância entre as categorias, porque teríamos um intervalo igual a si mesmo (ver *SPSS*, p. 5-6). Conseqüentemente, uma variável binária pode ser analisada com métodos apropriados para os níveis de medida nominal, ordinal ou intervalar.

2.3 Níveis de Medida e Testes Estatísticos

Os procedimentos estatísticos designados para um nível inferior de mensuração sempre podem ser utilizados com variáveis que apresentar um nível de medida mais alto (uma variável intervalar, por exemplo, pode ser analisada a partir de testes desenvolvidos para uma variável de nível nominal ou ordinal). Mas não se pode recorrer a procedimentos desenvolvidos para variáveis com um nível de medida mais alto para analisar variáveis com um nível inferior de medida (isto é, analisar uma variável nominal com os testes de uma variável intervalar).

A tabela 2.1 mostra tanto a relação entre os níveis de medida da primeira e da segunda variável (no caso de uma análise bivariada) e os procedimentos estatísticos apropriados, como o procedimento ou subprograma de SPSS que deve ser utilizado. Por exemplo, no caso onde a primeira variável é de nível nominal e a segunda de nível de dicotomia, o procedimento de SPSS para análise bivariada seria o CROSSTABS com testes de significância e de associação estatística da relação entre as duas variáveis com qui-quadrado, V de Cramer, etc.

Nos próximos itens descreveremos os quatro tipos de cartões de controle de um programa de SPSS:

(3.0) Cartões de Definição de Dados;
(4.0) Cartões de Criação e Modificação de Variáveis
(5.0) Cartões para Listagem e Modificação dos Arquivos de SPSS; e
(6.0) Cartões de Definição de Operações ou Procedimentos.

3. CARTÕES DE DEFINIÇÃO DE DADOS

Os cartões de definição de dados comunicam ao sistema os nomes e o número de variáveis (VARIABLE LIST), o número de observações ou casos (N OF CASES), o meio de entrada de dados (INPUT MEDIUM),

TABELA 2.1

ESTATÍSTICA DESCRITIVA E TESTES DE SIGNIFICÂNCIA E DE ASSOCIAÇÃO ESTATÍSTICA DA RELAÇÃO ENTRE DUAS VARIÁVEIS

NÍVEL DE MENSURAÇÃO DA PRIMEIRA VARIÁVEL	PROCEDIMENTOS COM UMA VARIÁVEL	PROCEDIMENTOS BIVARIADOS			
		Nível de mensuração da segunda variável			
		Dicotomia	Nominal	Ordinal	Intervalar ou razão
DICOTOMIA	Proporções, Percentagens, razões FREQUENCIES(1)	Diferença entre duas proporções Teste exato de Fisher, Q de Yale, Lambda Tau, etc. CROSSTABS(1)			
NOMINAL	Proporções, percentagens, razões FREQUENCIES(1)	Qui-quadrado V de Cramer Tau de Kendall Coeficiente de Incerteza D de Somer Eta, Lambda CROSSTABS(1)	Igual à nominal-dicotomia		
ORDINAL	Mediana, Quartis, decis	Mann-Whitney Smirnou, NPAR TESTS(1)	Análise de Variância com Ordenação NPAR TESTS*	Correlação "rank order" de Spearman ou de Kendall NONPAR CORR	
INTERVALAR OU RAZÃO	Média, Mediana, desvio padrão CONDESCRIPTIVE(1)	Diferença de Médias T-TEST(1)	Análise de Variância, F, Eta[a], BREAKDOWN(1)		Correlação de Regressão SCATTERGRAM PEARSON CORR(1) PARTIAL CORR(1)

FONTE — Baseado em uma tabela de Blalock, H.M. Blalock, *Social Statistics* (México: Mc Graw — Hill, 1972) e no manual de SPSS.

(1) Procedimento de SPSS.

e a localização de cada variável no cartão de dados (INPUT FORMAT), além de fornecerem outras informações para facilitar a leitura dos resultados.

3.1 O Formato geral de um Cartão de Controle SPSS (SPSS, p. 30)

Todos os cartões de SPSS têm um formato geral. Nas colunas 1 a 15 especifica-se o tipo de operação a ser executada. Nas colunas 16 a

80 especificam-se variáveis, rótulos, instruções, parâmetros, etc. que correspondem à operação definida nas colunas 1 a 15:

coluna

1	16	80
campo de definição do tipo de operação de SPSS	campo de especificação das instruções da operação	

Em se tratando de um arquivo gerado pelo usuário via terminal, observa-se que o número máximo de colunas, caso o usuário deseje que o arquivo seja numerado, é 72. Nesta situação, as colunas de 72 a 80 ficam reservadas para a numeração do arquivo. Além disso, se o arquivo for numerado, tem de se incluir o seguinte cartão no início do *deck*:

1	16
NUMBERED	YES

3.2 Os Cartões de Definição

3.2.1 RUN NAME — O Rótulo do Programa (SPSS, p. 72)

O RUN NAME é simplesmente o nome ou rótulo pelo qual o usuário quer identificar seu programa. É opcional e pode conter até 64 caracteres alfanuméricos.

1	16
RUN NAME	DADOS SOBRE AS TURMAS DE 76 77 78

3.2.2 VARIABLE LIST — A Lista das Variáveis (SPSS, p. 36-38)

Este cartão fornece a lista de variáveis que será lida nos cartões de dados. Outras variáveis podem ser desenvolvidas com base nestes dados originais, mediante os cartões de criação e modificação de variáveis (ver 4.0).

O nome da variável pode apresentar até oito (8) caracteres. O primeiro deve ser uma letra, mas os outros podem ser numéricos. A ordem dos nomes na VARIABLE LIST deve ser a mesma que a ordem em que as variáveis aparecem nos cartões de dados. Em nosso exemplo, as variáveis terão os seguintes nomes e ordem:

MAT, número de matrícula
SEXO, sexo do aluno
DIA, MES e ANO de nascimento

AREA, área de concentração
TUR, ano de entrada
ING, nota na prova de inglês
EST, nota na prova de estatística
RED, nota na prova de redação

Nosso cartão seria então:

1	16
VARIABLE LIST	MAT, SEXO, DIA, MES, ANO, AREA, TUR, ING, EST, RED

Também existe uma opção para definição de variáveis, sob forma de uma enumeração, que é bastante útil quando se quer definir muitas variáveis. Em nosso exemplo de dez variáveis, o formato seria:

1	16
VARIABLE LIST	VAR001 TO VAR010

Este cartão significa: lê variáveis 1 a (*to*) 10.

3.2.3 INPUT MEDIUM — Meio de Entrada dos Dados (SPSS, p. 39-40)

O cartão INPUT MEDIUM informa ao sistema onde ele deve ler os dados. Temos três (3) opções:

CARD — cartões
TAPE — fita magnética
DISK — disco magnético

Em nosso exemplo, vamos utilizar cartões

1	16
INPUT MEDIUM	CARD

A utilização de fitas ou discos exige a colocação de cartões de JCL que fornecem ao sistema e ao operador as informações necessárias sobre a fita a ser montada ou o arquivo a ser lido.

3.2.4 N OF CASES — Número de Casos (SPSS, p. 40-41)

Este cartão informa ao sistema sobre o número de casos. O caso é a unidade de análise, como, por exemplo, alunos, escolas, cidades ou estados. Em nosso exemplo, temos uma amostra de 90 alunos:

1	16
N OF CASES	90

No caso de utilizar fitas ou discos, em não se sabendo o número exato de casos, pode-se colocar UNKNOWN (desconhecido).

```
1                    16
N OF CASES           UNKNOWN
```

3.2.5 *INPUT FORMAT — Formato de Entrada dos Dados (SPSS, p. 41-46)*

Este cartão informa ao sistema a posição de cada variável nos cartões de dados. O INPUT FORMAT é semelhante ao FORMAT em FORTRAN. Há dois tipos de INPUT FORMAT: FIXED e FREEFIELD; mas, por ora, vamos somente utilizar o FIXED. FIXED significa que a variável vai sempre ficar entre as mesmas colunas do cartão (o mesmo campo), vai ser sempre alinhada pela direita. Ver qualquer texto sobre FORTRAN IV para maiores detalhes.

Há três (3) tipos de variáveis que podem ser utilizadas:

A — para letras (esta não será aqui utilizada).

F — para variáveis numéricas. Estas podem ser números inteiros (*integers*) ou decimais.

X — para saltar colunas.

Há três (3) parâmetros que devem ser definidos para a variável F: nFw.d

n — número de variáveis com parâmetros *w* e *d*.

w — largura do campo em colunas.

d — número de dígitos à direita do ponto (decimais)

Por exemplo, a expressão 3F5.0 avisa ao computador que existem três (3) campos de dados, com cinco (5) colunas em cada um, e cada variável tem ZERO (0) lugares depois do ponto.

O *FORMAT* de nosso exemplo:

```
1                    16
INPUT FORMAT         FIXED (F6.0, F2.0, 3F3.0, F2.0, 4F3.0)
```

O FORMAT controla a localização do ponto decimal, daí decorrendo a necessidade de defini-lo com todo o cuidado. A tabela 3.1 mostra o formato, o valor no cartão de dados e o valor lido pelo computador. Se o valor no cartão tem um ponto, ele tem prioridade sobre o FORMAT, ou seja, é o ponto e não o FORMAT que determina o valor lido e arquivado.

TABELA 3.1

O FORMATO E OS VALORES LIDOS

FORMATO	VALOR NO CARTÃO (Colunas)	VALOR LIDO E ARQUIVADO
	123	
F3.0	100	100
F3.1	100	10.0
F3.2	100	1.00
F3.1	1.0	1.0
F3.2	.10	0.10
F3.3	10.	10.0

O elemento *nX* permite que se saltem *n* colunas no cartão de dados. Por exemplo, 3X quer dizer saltar três (3) colunas. A barra (/) informa ao sistema para saltar um cartão. O cartão seguinte manda o sistema ler três (3) cartões de dados, apresentando cada um deles três variáveis:

```
1                16
INPUT FORMAT     FIXED (F10.0, 5X, 2F7.3/3F8.0/3F10.3)
```

3.2.6 VAR LABELS — Designações das Variáveis (SPSS, p. 62-63)

Este cartão de controle permite a melhor identificação das variáveis através de designações (rótulos) de até 40 caracteres alfanuméricos através de outros símbolos válidos de FORTRAN com exceção da barra (/):

```
1                16
VAR LABELS       MAT, NUMERO DA MATRICULA
VAR LABELS       SEXO, SEXO DO ALUNO
```

Em vez de repetir VAR LABELS para cada variável, poderíamos utilizar uma barra (/) para continuar o primeiro cartão

```
1                16
VAR LABELS       MAT, NUMERO DA MATRICULA/
                 SEXO, SEXO DO ALUNO/
                 RED, NOTA DE REDACAO
```

Observe-se que o último cartão *não tem* uma barra (/) porque não é continuação da mesma operação.

O VAR LABELS é opcional. Para colocação no *deck,* ver 3.5.

3.2.7 VALUE LABELS — Codificação das Variáveis (SPSS, p. 59-61)

Com o cartão VALUE LABELS pode-se mandar o sistema imprimir uma designação ou rótulo para cada categoria de uma variável não-contínua. Cada uma destas designações pode ter até 20 (vinte) caracteres alfanuméricos e pode ser continuada com uma barra (/), como no caso do VAR LABELS e vários outros cartões de SPSS. Este cartão é opcional. Em nosso exemplo poderíamos empregar VALUE LABELS não só para as três variáveis não-contínuas como também para as novas variáveis não-contínuas geradas (ver parte 4.0):

```
1                16
VALUE LABELS     SEXO (1) MASCULINO (2) FEMININO/AREA(1)
                 PLANEJAMENTO (2) METODOS E TECNICAS (3)
                 ACONSELHAMENTO/TUR (1) TURMA DE 76 (2)
                 TURMA DE 77 (3) TURMA DE 78
```

Pode-se observar que a barra (/) tem três funções em SPSS: a) saltar cartões, no cartão de INPUT FORMAT; b) continuação, neste e noutros tipos de cartões de controle; c) divisão de variáveis, nos cartões de modificação de dados (ver parte 4.0).

3.2.8 MISSING VALUES — Valores Excluídos (SPSS, p. 57-59)

Em pesquisas em ciências sociais há situações onde não existem dados para todas as variáveis de um caso, ou onde o pesquisador quer excluir alguns valores dos cálculos. Por exemplo, existem casos em que a pessoa não quer responder à pergunta ou em que a pergunta não se aplica. Podemos definir um valor para todas as situações onde nos faltam dados. Supondo que 99 é definido como este valor, todos os casos onde faltam dados seriam codificados 99. Em nosso exemplo, o cartão MISSING VALUES seria:

```
1                16
MISSING VALUES   SEXO TO RED (99)
```

Poderíamos também eliminar casos com certas características dos cálculos. Supondo, por exemplo, que queremos fazer cálculos incluindo

somente o sexo masculino e a área de planejamento, nosso cartão de MISSING VALUES corresponderia a:

```
1                    16
MISSING VALUES       SEXO (2)/AREA (2,3)
```

Note-se então o uso da barra (/) para continuação e o da vírgula para inclusão de mais de um *missing value*.

3.3 READ INPUT DATA — ler os cartões de dados

Informa ao sistema para ler os dados. Este cartão aparece depois que se opera com o primeiro procedimento estatístico (ver 3.5 para a ordem do *deck*).

3.4 Modificação Permanente e Temporária (com o asterisco)

Temos a opção de modificar uma variável para todas as operações ou subprogramas (sem asterisco) ou somente para uma operação (com asterisco). Nos casos do *COMPUTE, *RECODE, *IF, *SELECT IF, a modificação da variável se presta somente para a operação ou subprograma que os segue e não para outras operações. Por exemplo, na seqüência de cartões abaixo, o *IF e o *COMPUTE, se prestariam somente para o subprograma REGRESSION e não para o BREAKDOWN:

```
1                    16
*IF                  (SEXO EQ 2) SEXO=0
*COMPUTE             IDSEX=IDADE*SEXO
REGRESSION           VARIABLES=EST, IDADE, SEXO, IDSEX/
/STATISTICS/           REGRESSION=EST WITH IDADE
/BREAKDOWN             SEXO IDSEX (1)
                     ALL
                     TABLES=EST BY SEXO BY TUR
```

3.5 FINISH — fim

Informa ao sistema que o programa de SPSS está terminado. Aparece como o último cartão de controle de SPSS.

4. OS CARTÕES DE CRIAÇÃO E MODIFICAÇÃO DE VARIÁVEIS

Estes cartões permitem a geração, transformação e redefinição de variáveis.

4.1 COMPUTE — Cálculos Aritméticos (SPSS pp. 96-101)

O cartão de controle COMPUTE determina não só a transformação de variáveis com funções aritméticas e algumas funções especiais como também a geração de novas variáveis. A tabela 4.1 mostra as operações que podem ser feitas com o cartão COMPUTE.

TABELA 4.1

OPERAÇÕES ARITMÉTICAS E OUTRAS FUNÇÕES POSSÍVEIS COM O CARTÃO COMPUTE

SÍMBOLO	OPERAÇÃO	EXEMPLO
/	Divisão	X = A/B
*	Multiplicação	X = A*B
+	Adição	X = A + B
—	Subtração	X = A — B
**	Exponenciação	X = A**B
SQRT	Raiz Quadrada	X = SQRT(A)
LN	Logaritmo Natural	X = LN(A)
LG10	Logaritmo Base 10	X = LG10(A)
EXP	Exponencial (e^{arg})	X = EXP(A + C)

Assim sendo, poderíamos em nosso exemplo calcular a idade do aluno em 31-12-79, levando-se em consideração a data de nascimento:

```
1                    16
COMPUTE              IDADE = 1979 — ANO
```

Também seria possível gerar uma nova variável (IDCAT) equivalente a IDADE:

```
1                    16
COMPUTE              IDCAT = IDADE
```

Vamos transformar IDCAT em uma variável não-contínua ou categorial com o cartão RECODE (ver parte 4.2). Teremos, então, duas formas da variável idade: a contínua (IDADE) e a categorial (IDCAT).

Existe a opção de utilização de parênteses na geração de funções mais complexas:

```
1                    16
COMPUTE              X = (((P-Z)*(PO/VAR500)) — 1.0)**3
```

4.2 RECODE — Recodificação de valores (SPSS p. 90-96)

Este cartão permite ao usuário converter uma variável contínua em uma variável categorial ou não-contínua. O seguinte cartão transforma a variável IDCAT (que é igual a IDADE) em uma variável categorial:

```
1                    16
RECODE               IDCAT (LOWEST THRU 24=1) (25 THRU 29=2)
                     (30 THRU 34=3) (35 THRU 39=4) (40
                     THRU HIGHEST = 5)
```

Os grupos de idade gerados com este cartão serão:

(1) Até 24 anos
(2) 25 — 29 anos
(3) 30 — 34 anos
(4) 34 — 39 anos
(5) 40 ou mais anos

LOWEST significa mais baixo e THRU (*through*) quer dizer inclusive. Por isso, todos os valores menores que ou iguais a 24 no cartão acima são recodificados no primeiro grupo (1). HIGHEST significa mais alto e todos os valores iguais ou superiores a 40 estarão classificados no grupo 5.

Pode-se utilizar a convenção "TO" para recodificar uma série de variáveis da mesma maneira:

```
1                    16
RECODE               VAR001 TO VAR100 (1 THRU 5=1)
                     (6 THRU 9=2) (10 THRU HIGHEST=3)
```

Se as variáveis contínuas têm decimais, deve-se fazer uma superposição dos limites dos grupos:

```
1                    16
RECODE               X, Y, Z (0.10 THRU 0.25=1) (0.25 THRU
                     0.50=2) (0.50 THRU HIGHEST=3)
```

Ver tabela 7.1 para a colocação do RECODE dentro do Deck de SPSS.

4.3 IF, SELECT IF — Expressões Lógicas (SPSS, p. 101-107, 128)

Às vezes precisa-se modificar uma variável ou gerar outras em função de uma expressão lógica. A tabela 4.2 mostra as diferentes expressões lógicas presentes no SPSS.

TABELA 4.2

EXPRESSÕES LÓGICAS DE SPSS

EXPRESSÃO		SÍMBOLO
Português	Inglês	
MAIOR QUE OU IGUAL A	GREATER THAN OR EQUAL TO	GE
MENOR QUE OU IGUAL A	LESS THAN OR EQUAL TO	LE
MAIOR QUE	GREATER THAN	GT
MENOR QUE	LESS THAN	LT
IGUAL A	EQUAL TO	EQ
NÃO IGUAL A	NOT EQUAL TO	NE
E	AND	AND
OU	OR	OR

Por exemplo, suponhamos que se deseje saber a idade da pessoa em 31-07-78 ao invés de 31-12-79. Neste caso, todas as pessoas que fazem anos depois de julho (7) teriam um ano a menos, como mostra o exemplo na parte 4.1. Então teríamos que subtrair um ano desta idade para todos estes alunos, com o seguinte cartão:

```
1                16
IF               (MES GE 8) IDADE=IDADE—1
```

Este cartão diz "Se o MES é maior que ou igual a 8, IDADE é igual a IDADE menos um (1)".

Podemos também definir todos os outros valores da variável X igual a 999, quando a variável Y for igual a 2:

```
1                16
IF               (Y EQ 2) X = 999
```

Pode-se combinar expressões lógicas com as palavras AND (e) e OR (ou). Por exemplo, se quiséssemos multiplicar VAR001 por 2, quando X for igual a 1 ou quando Z for maior que ou igual a 3, escreveríamos o seguinte cartão de controle:

```
1                16
IF               (X EQ 1 OR Z GE 3) VAR001=VAR001*2
```

Um exemplo do uso de AND seria: se S for menor que 3 e P for menor que ou igual a 4, Z é igual à raiz quadrada de VAR555:

```
1                    16
IF                   (S LT 3 AND P LE 4) Z=SQRT (VAR555)
```

Deve-se ter cuidado com os parênteses quando se utiliza mais de um AND ou OR com a mesma variável. Por exemplo, os dois cartões abaixo são corretos, mas significam coisas diferentes:

```
1                    16
IF                   (A EQ 9 OR (A GT O AND A LT 5)) A=1
IF                   (A EQ 9 OR A GT O GT O AND A LT 5) A =1
```

A diferença é que há uma separação entre as expressões lógicas no primeiro cartão, o que não ocorre no segundo. Então, o primeiro cartão significa: coloque A igual a 1, se A for igual a 9 ou maior que zero (0) e menor que 5. O segundo cartão informa: coloque A igual a 1, se A for igual a 9 e *menor que 5,* ou maior que zero (0) e menor que 5.

Operações aritméticas também podem ser indicadas dentro da expressão lógica. Por exemplo, se o produto de VAR022 multiplicado pela VAR333 é maior que VAR555, VAR888 é igual ao logaritmo de base 10 de VAR999.

```
1                    16
IF                   (VAR022 * VAR333 GT VAR555)
                     VAR888=LG10 (VAR999)
```

Usa-se SELECT IF (escolha se) para se escolher uma parte da população segundo um ou mais critérios. Assim, se quiséssemos processar somente os casos em que se assinala RENDA superior a Cr$ 1.500,00, bastaria colocar o cartão:

```
1                    16
SELECT IF            (RENDA GT 1500)
```

Então, só os casos com RENDA Superior a 1.500 entrariam nos cálculos. Pode-se utilizar o AND e o OR para se formularem com o SELECT IF expressões mais complicadas. Por exemplo, o cartão seguinte exclui dos cálculos todos os casos com RENDA inferior a 1.500 e de SEXO com valor 2:

```
1                    16
SELECT IF            (RENDA GT 1500 AND SEXO NE 2)
```

Todos os símbolos do IF descritos na tabela 4.2 podem também ser usados com o SELECT IF.

4.4 DO REPEAT — Operações Repetidas (SPSS, p. 121-125)

Às vezes o pesquisador quer fazer o mesmo cálculo com uma série de variáveis. Por exemplo, pode-se querer dividir VAR001 até VAR050 por população total (POP). Ao invés de se preparar um cartão de COMPUTE para cada uma destas 50 variáveis, podemos repetir a operação 50 vezes com o cartão DO REPEAT. Em SPSS deve-se designar por variável macro aquela que representa todas as variáveis que vão ser mudadas (neste caso VAR001 até VAR050). Os seguintes cartões resulta da divisão de todas estas 50 variáveis por POP:

```
1                   16
DO REPEAT           XXX = VAR001 TO VAR050
COMPUTE             XXX = XXX/POP
END REPEAT
```

A variável XXX (a variável macro) é "conhecida" pelo sistema somente durante a operação do DO REPEAT. Em outras palavras, o sistema usa esta variável somente até o END REPEAT aparecer. O END REPEAT marca o fim da operação. O rótulo de variável macro não pode ser utilizado em nenhuma outra parte do Programa de SPSS.

Os cartões de controle que podem aparecer dentro do DO são:

```
IF                  SELECT IF
COMPUTE             MISSING VALUES
COUNT               RECODE
```

Forneçamos um exemplo um pouco mais sofisticado. Vamos recorrer ao mesmo COMPUTE supracitado; no entanto, vamos gerar mais 50 variáveis, de VAR301 até VAR351. Teremos então uma outra variável macro, YYY:

```
1                   16
DO REPEAT           XXX = VAR001 TO VAR050/
                    YYY = VAR301 TO VAR350/
COMPUTE             YYY = XXX
COMPUTE             YYY = YYY/POP
RECODE              YYY (.00 THRU .100=1) (.100 THRU
/IF                 .500=2) (.500 THRU HIGHEST=3)
/MISSING VALUES     (YYY EQ 3) YYY = 999
/END REPEAT         YYY (999)
```

O primeiro COMPUTE iguala as variáveis VAR001 até VAR050 às variáveis VAR301 até VAR350 (ou VAR001 = VAR301, VAR002 = = VAR302, VAR003 = VAR303 etc.). O segundo divide o total das variáveis (definido pela variável macro YYY) pela variável POP (ou VAR301/POP, VAR302/POP etc.). O RECODE reúne os valores da variável macro YYY em grupos. O IF substitui todos os valores da variável macro YYY por 999 se o valor inicial desta for 3. Este valor 999 é definido como um MISSING VALUE. Então, todos os valores 999 da variável macro vão ser eliminados dos cálculos.

5. CARTÕES PARA LISTAGEM E MODIFICAÇÃO DOS ARQUIVOS DE SPSS

O sistema de informação de SPSS é extenso e flexível; aqui, ao examinarmos o LIST CASES, vamos focalizar somente um dos comandos mais comumente utilizados (ver SPSS, p. 133-175 para os restantes).

5.1 LIST CASES — A listagem dos casos (SPSS, p. 137-139)

O usuário deve sempre verificar os dados para ver se houve erros de perfuração ou de formato. O cartão de controle LIST CASES permite a listagem dos valores das variáveis de cada caso. O formato geral deste cartão é:

1	16
LIST CASES	CASES = número/VARIABLES = lista de variáveis

Com o cartão que se segue abaixo poderíamos listar as variáveis SEXO, IDADE, TUR para 90 casos.

1	16
LIST CASES	CASES = 90/VARIABLES = SEXO, IDADE, TUR

Ou todas as variáveis:

1	16
LIST CASES	CASES = 90/VARIABLES = ALL

Existem dois tipos de formato para o LIST CASES: o normal e o condensado. O formato condensado é utilizado automaticamente para todas as situações quando são especificadas menos de 12 variáveis, e o normal no caso de 12 ou mais variáveis.

O cartão LIST CASES deve aparecer *antes* de um cartão de procedimento estatístico (CONDESCRIPTIVE, CROSSTABS, etc.). Por exemplo,

1	16
LIST CASES	CASES = 34/VARIABLES = SEXO,
	ANO, IDADE,
CONDESCRIPTIVE	ING, RED, EST
STATISTICS	ALL

Este último cartão gerou a listagem do exemplo 5.1.

EXEMPLO 5.1

LIST CASES: LISTAGEM DOS VALORES DAS VARIÁVEIS SEXO, ANO E IDADE (FORMATO CONCLUSADO)

CASE-N	SEXO	ANO	IDADE
1	1.	43.	36.
2	1.	40.	39.
3	1.	52.	27.
4	1.	37.	42.
5	1.	39.	40.
6	2.	39.	40.
7	2.	41.	38.
8	2.	49.	30.
9	1.	45.	34.
10	1.	42.	37.
11	1.	36.	43.
12	2.	38.	44.
13	2.	41.	38.
14	2.	30.	49.
15	2.	45.	34.
16	2.	52.	27.
17	2.	41.	38.
18	2.	46.	33.
19	2.	48.	31.
20	1.	41.	38.
21	2.	40.	39.
22	2.	40.	39.
23	2.	39.	40.
24	2.	37.	42.
25	2.	45.	34.
26	2.	37.	42.
27	2.	50.	29.
28	2.	42.	37.
29	2.	40.	39.
30	2.	46.	33.
31	1.	28.	51.
32	2.	41.	38.
33	2.	41.	38.
34	2.	39.	40.

6. CARTÕES DE DEFINIÇÃO DE OPERAÇÕES OU PROCEDIMENTOS

6.1 CONDESCRIPTIVE — Estatísticas Descritivas para Variáveis Contínuas (SPSS, p. 181-193)

Este subprograma gera estatísticas descritivas para variáveis contínuas. O exemplo 6.1 mostra a listagem (*output*) para os seguintes cartões de controle

1	16
CONDESCRIPTIVE	RED, ING, EST
STATISTICS	ALL

EXEMPLO 6.1

CONDESCRIPTIVE: ESTATÍSTICAS DESCRITIVAS PARA AS VARIÁVEIS CONTÍNUAS, RED, EST E ING

```
VARIABLE  RED        NOTA DE REDACAO

MEAN              7.233      STD ERROR        0.156      STD DEV          1.477
VARIANCE          2.181      KURTOSIS        -0.196      SKEWNESS        -0.671
RANGE             6.000      MINIMUM          4.000      MAXIMUM         10.000
SUM             651.000

VALID OBSERVATIONS -      90          MISSING OBSERVATIONS -      0

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

VARIABLE  EST        NOTA DE ESTATISTICA

MEAN              7.100      STD ERROR        0.166      STD DEV          1.573
VARIANCE          2.473      KURTOSIS        -0.490      SKEWNESS         0.044
RANGE             6.000      MINIMUM          4.000      MAXIMUM         10.000
SUM             639.000

VALID OBSERVATIONS -      90          MISSING OBSERVATIONS -      0

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

VARIABLE  ING        NOTA DE INGLES

MEAN              8.144      STD ERROR        0.169      STD DEV          1.604
VARIANCE          2.574      KURTOSIS        -0.372      SKEWNESS        -0.558
RANGE             6.000      MINIMUM          4.000      MAXIMUM         10.000
SUM             733.000

VALID OBSERVATIONS -      90          MISSING OBSERVATIONS -      0
```

Aqui o usuário está solicitando todas as estatísticas descritivas (ALL significa todas, em inglês). A tabela 6.1 apresenta o número de identificação da estatística no SPSS, seu nome inglês e português e referências ao livro de Spiegel e ao manual de SPSS:

TABELA 6.1

AS ESTATÍSTICAS DESCRITIVAS DO CONDESCRIPTIVE

NÚMERO	DESIGNAÇÃO USADA EM ESTATÍSTICA		OBRAS DE REFERÊNCIA (n.º da página)	
	Em inglês	Em português	Estatística (Murray R. Spiegel)	Manual de SPSS
1	Mean	Média	71—74	183—184
2	Standard Error	Erro-Padrão	238	184
5	Standard Deviation	Desvio-Padrão	113—116	184
6	Variance	Variância	112—113	184
7	Kurtosis	Curtose	147	185
8	Skewness	Assimetria	145	184—185
9	Range	Amplitude Total	108	182
10	Minimum	Mínimo		182
11	Maximum	Máximo		182

Suponhamos que o usuário deseje apenas a média e o desvio-padrão; o cartão seria então:

1	16
STATISTICS	1,5

Como a média, o erro-padrão e o desvio-padrão são discutidos na maioria dos textos de Estatística básica (ver, por exemplo, as páginas citadas na tabela 6.1), não será necessário tratá-los aqui. Contudo, focalizaremos rapidamente a assimetria e a curtose, uma vez que seu cálculo no SPSS é um pouco diferente do que aparece na maioria destes textos.

Assimetria (*skewness*) é o grau de desvio de uma distribuição em relação a uma distribuição normal. Embora existam vários métodos para o cálculo de assimetria de uma distribuição, o SPSS usa o terceiro momento (SPSS, p. 184-185). A figura 6.1 mostra como interpretar os resultados. Em nosso exemplo, as notas de estatística têm uma distribuição quase simétrica ou normal (0,044), enquanto as distribuições das notas de inglês e redação se mostram concentradas à direita (0,558 e 0,671), ou seja, em assimetria negativa.

ASSIMETRIA ("SKEWNESS")

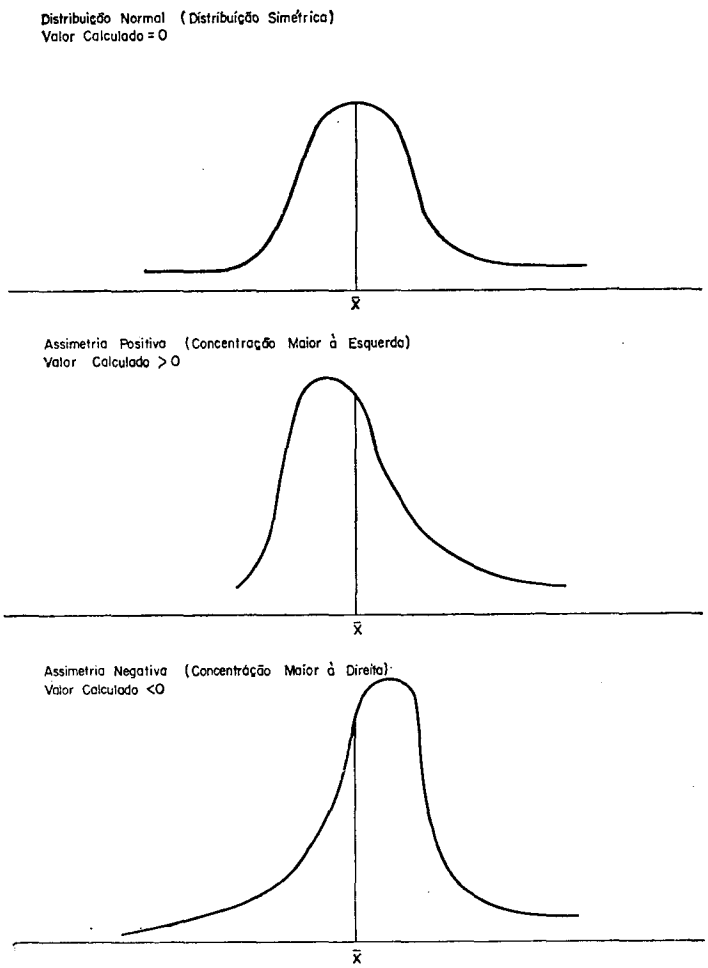


Fig. 6.1

CURTOSE ("KURTOSIS")

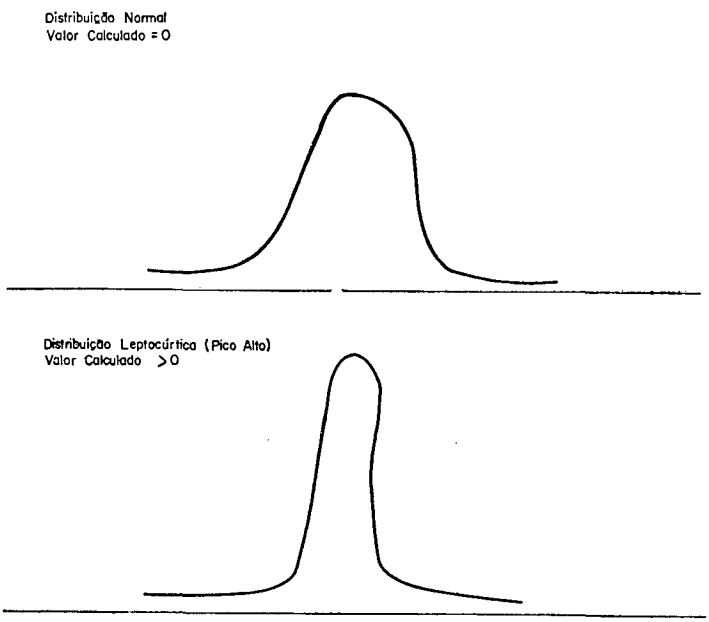


Fig. 6.2

Curtose (*kurtosis*) designa o grau de achatamento de uma distribuição em relação a uma curva normal (Spiegel, p. 147). Como no SPSS o valor do coeficiente do momento de curtose de uma curva normal (que é igual a 3) é subtraído do coeficiente calculado, este coeficiente é igual a zero quando a distribuição é normal (ver figura 6.2); positiva quando a distribuição tem um pico relativamente alto (leptocúrtica); e negativa quando ela é uma distribuição relativamente plana (platicúrtica). Em nosso exemplo as distribuições das três notas são ligeiramente mais planas do que a distribuição normal (platicúrticas): RED = = 0,196, EST = 0,490 e ING = 0,372.

6.2 FREQUENCIES — Distribuição de Freqüência e Estatísticas Descritivas para Variáveis Categoriais. (SPSS, p. 194-202)

O subprograma FREQUENCIES gera distribuições de freqüência de uma dimensão, estatísticas descritivas e histogramas para variáveis categorias (não-contínuas). Temos dois modos de operar:

INTEGER (quando os valores das variáveis correspondem sempre a números inteiros)

GENERAL (quando os valores das variáveis podem ser ou não números inteiros)

O GENERAL exige menos trabalho na preparação de cartões, mas é menos eficiente em termos de tempo de computação. Já o INTEGER exige mais trabalho na preparação dos cartões, mas é muito mais eficiente que o GENERAL em relação a tempo de computação e usa menos espaço na memória. A escolha do modo de operar deve ser feita em função do número de casos e do custo de tempo e espaço na memória do computador. Em caso de programas grandes, a poupança de custo de computação deve compensar o esforço na preparação dos cartões.

O formato geral do FREQUENCIES com o modo de operar GENERAL é:

```
1                    16
FREQUENCIES          GENERAL = lista de variáveis
```

Por exemplo, os seguintes cartões geram resultados para três variáveis, inclusive a nota de estatística (ver exemplo 6.2):

```
1                    16
FREQUENCIES          GENERAL = IDCAT, SEXO, EST
STATISTICS           ALL
OPTIONS              8
```

EXEMPLO 6.2

FREQUENCIES: DISTRIBUIÇÃO DE FREQÜÊNCIA, ESTATÍSTICAS PARA A VARIÁVEL

```
AREA      AREA DE CONCEN

                                                    RELATIVE    ADJUSTED      CUM
                                      ABSOLUTE        FREQ        FREQ        FREQ
CATEGORY LABEL                 CODE     FREQ         (PCT)       (PCT)       (PCT)

PLANEJ                          1.       43          47.8        47.8        47.8
MET E TEC                       2.       34          37.8        37.8        85.6
ACONS                           3.       13          14.4        14.4       100.0
                                       ------      ------      ------
                               TOTAL     90         100.0       100.0
```

```
AREA      AREA DE CONCEN

     CODE
          I
       1. ****************************************** (     43)
          I  PLANEJ
          I
          I
       2. ********************************** (     34)
          I  MET E TEC
          I
          I
       3. ************* (     13)
          I  ACONS
          I
          I.........I.........I.........I.........I.........I
          0         10        20        30        40        50
          FREQUENCY
```

MEAN	1.667	STD ERR	0.076	MEDIAN	1.559
MODE	1.000	STD DEV.	0.719	VARIANCE	0.517
KURTOSIS	-0.854	SKEWNESS	0.598	RANGE	2.000
MINIMUM	1.000	MAXIMUM	3.000		

VALID CASES 90 MISSING CASES

O formato geral do FREQUENCIES com o modo INTEGER é:

```
1                16
FREQUENCIES INTEGER = variável
                      ou lista (valor mínimo, valor máximo)
                      de var.

                    variável (valor mínimo, valor máximo)
                    ou lista
                    de var.
```

O cartão que seria utilizado para análise das mesmas variáveis especificadas no cartão anterior é

```
1                16
FREQUENCIES      INTEGER = IDCAT (1,5) SEXO (1,2)
STATISTICS       ALL
OPTIONS          8
```

O programa FREQUENCIES gera as estatísticas descritivas iguais ao CONDESCRIPTIVE (com a mesma numeração), mais:

3. Mediana (Median) Ver Spiegel, p. 74

4. Moda (Mode) Ver Spiegel, p. 74-75

A opção 8 no cartão OPTIONS gera histogramas das distribuições. Existem outras opções para controle da listagem e dos MISSING VALUES. Sem o cartão OPTIONS, estes MISSING VALUES são excluídos dos cálculos. Ver SPSS (p. 200-201) para um estudo mais detalhado destas opções. O número máximo de variáveis que pode aparecer no cartão FREQUENCIES é 500. Para outras limitações, ver SPSS, p. 201-202.

6.3 CROSSTABS — Tabulações Cruzadas (SPSS, p. 218-248)

Depois de analisar as distribuições das variáveis categoriais mediante o subprograma FREQUENCIES o pesquisador normalmente quer fazer tabulações cruzadas, às vezes aplicando testes estatísticos de associação e significância estatística entre as variáveis tabuladas.

Existem dois modos de fazer tabulações cruzadas com o cartão CROSSTABS: GENERAL (Geral) e INTEGER (Número Inteiro). Para

uma análise das vantagens e desvantagens de cada modo, ver FREQUENCIES (6.2). O formato geral do cartão com o modo GENERAL é:

```
1                16
CROSSTABS        TABLES = variável BY  variável      BY...BY
                          ou lista    ou lista
                          de var.     de var.

                 variável
                 ou lista
                 de var.
```

Por exemplo, o cartão que se segue produz a tabela apresentada no exemplo 6.3:

```
1                16
CROSSTABS        TABLES = EST BY SEXO
```

"BY" significa "por" em português e o cartão pode ser lido com uma tabela de nota estatística por sexo. A convenção, "TO", também pode ser utilizada. Observa-se que as duas variáveis devem ser categoriais. Uma variável pode ser cruzada com mais de uma variável, com um cartão (SEXO por IDCAT e SEXO por TUR):

```
1                16
CROSSTABS        TABLES = SEXO BY IDCAT, TUR
```

Pode-se também produzir tabelas com mais de duas (2) dimensões, como no exemplo do cartão abaixo que geraria uma tabela SEXO por (BY) IDADE para cada valor de TUR (76, 77, 78):

```
1                16
CROSSTABS        TABLES = SEXO BY IDCAT BY TUR
```

Pode-se proceder a estas duas operações com um cartão, utilizando a barra (/) para continuação:

```
1                16
CROSSTABS        TABLES = SEXO BY IDCAT SEXO BY
                   IDCAT BY IDCAT
```

De acordo com o modo INTEGER o usuário deve incluir um cartão VARIABLES antes do cartão TABLES. Este cartão VARIABLES mostra

EXEMPLO 6.3

CROSSTABS: TABULAÇÃO CRUZADA DAS VARIÁVEIS SEXO E NOTA NA PROVA DE ESTATÍSTICA (EST) COM CÁLCULO DO QUI-QUADRADO E V DE CRAMER

```
* * * * * * * * * * * * * * * * * * * C R O S S T A B U L A T I O N   O F * * * * * * * * *
   SEXO      SEXO DO ALUNO                                  BY  EST      NOTA DE ESTATISTICA
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
```

SEXO	EST	ATE 6	7.	8.	9 +	ROW TOTAL
MASCULINO	1.	11	7	11	7	36
		30.6	19.4	30.6	19.4	40.0
		33.3	33.3	55.0	43.8	
		12.2	7.8	12.2	7.8	
FEMININO	2.	22	14	9	9	54
		40.7	25.9	16.7	16.7	60.0
		66.7	66.7	45.0	56.3	
		24.4	15.6	10.0	10.0	
COLUMN TOTAL		33	21	20	16	90
		36.7	23.3	22.2	17.8	100.0

CHI SQUARE = 2.96875 WITH 3 DEGREES OF FREEDOM SIGNIFICANCE = 0.3965
CRAMER'S V = 0.18162

os valores máximos de cada variável utilizada no cartão TABLES e tem o seguinte formato geral:

```
1                16
CROSSTABS        VARIABLES = variável (valor mínimo, valor máximo)
                             ou lista
                             de var.
                 variável (valor mínimo, valor máximo)
                 ou lista
                 de var.
```

No modo INTEGER, o exemplo acima seria codificado:

```
1                16
CROSSTABS        VARIABLES = SEXO (1,2) IDCAT (1,5)
                   TUR (76, 78)
                 TABLES = SEXO BY IDCAT SEXO BY
                   IDCAT BY TUR
                 STATISTICS ALL
```

A tabela 6.2 mostra as estatísticas de significância e de associação da relação entre duas variáveis geradas por *SPSS*, (p. 242-243). Já a tabela 6.3 ilustra o tipo de teste que deve ser aplicado com diferentes níveis de mensuração da primeira e da segunda variável.

TABELA 6.2

AS ESTATÍSTICAS DE SIGNIFICÂNCIA E ASSOCIAÇÃO ENTRE DUAS VARIÁVEIS DO SUBPROGRAMA CROSSTABS

NÚMERO	DESIGNAÇÃO USADA EM ESTATÍSTICA		OBRAS DE REFERÊNCIA (n.º da página)	
	Em inglês	Em português	Estatística (Murray R. Spiegel)	Manual de SPSS
1	Chi-squarea	Qui-quadrado	p.331—337	p.223—224
2	Phi for 2×2 tables Cramer's V	Phi para tabelas de 2×2 V de Cramer para tabelas maiores que 2×2		224
3	Contingency coefficient	Coeficiente de contingência	337	225
4	Lambda[b]	Lambda[b]		225—226
5	Uncertainty	Coeficiente de incerteza		226—227
6	Kendall's Tau b	Tau de Kendall b (tabela quadrada)		227—228
7	Kendall's Tau c	Tau de Kendall c (tabela retangular)		228
8	Gamma	Gama		228
9	Sommer's D[b]	D de Sommer[b]		229
10	Eta	Eta	230	230

a. Prova exata de Fisher no caso de tabelas 2 × 2 e/ou com menos de 21 casos.
b. simétrica e assimétrica.

TABELA 6.3

OS PROCEDIMENTOS ESTATÍSTICOS DO CROSSTABS, SEGUNDO OS NÍVEIS DE MENSURAÇÃO DA PRIMEIRA E DA SEGUNDA VARIÁVEL

PROCEDIMENTO ESTATÍSTICO	NÍVEL DE MENSURAÇÃO	
	1.ª Variável	2.ª Variável
Qui-quadrado V de Cramer Coeficiente Contigência Lambda	Nominal	Nominal
Tau b Tau c Gama O de Somer	Ordinal	Ordinal
Eta	Intervalo	Nominal

O qui-quadrado é utilizado freqüentemente para testar hipóteses sobre a independência de duas variáveis de nível de mensuração nominal (Hoel, p. 275-291). Faremos aqui o teste da hipótese nula de que a nota de Estatística (EST) e o sexo do aluno (SEXO) apresentados na tabela de contingência no exemplo 6.3 são independentes, ou seja, de que o desempenho na prova independe do sexo do aluno. Lembre-se, como foi discutido anteriormente, que as notas das três provas têm um nível de medida superior ao nominal (ver 2.2.3), mas resolvemos aqui utilizar um teste de um nível de medida inferior. Foi preciso agrupar um pouco as categorias de EST para atingir o mínimo de 5 observações esperadas em cada célula da tabela, utilizando-se para este fim o cartão RECODE (ver RECODE no exemplo 7.1).

A primeira etapa neste teste de hipótese é escolher-se o nível de significância, ou seja, a probabilidade de se cometer um erro do tipo I (rejeitar uma hipótese quando ela deve ser aceita). Como é tradicional neste tipo de pesquisa em ciências sociais, adotamos o nível de significância de 0,05. Com este nível a hipótese seria rejeitada quando deve ser aceita no máximo de 5 em cada 100 tentativas (para discussões do conceito de significância estatística ver Hoel, p. 201-238, Rodrigues, p. 85-105 ou Spiegel, p. 276-309).

Para chegarmos ao valor crítico do qui-quadrado com este nível de significância (0,50) precisamos calcular o número de graus de liberdade ("DEGREES OF FREEDOM") que é obtido com base no emprego da seguinte equação:

$$gl = (\text{N.º de linhas} - 1)\ (\text{N.º de colunas} - 1)$$

Com seus 3 graus de liberdade, o valor crítico do qui-quadrado em nosso exemplo seria 7,817. Visto que o valor calculado do qui-quadrado era somente 2,968, *não* poderemos rejeitar a hipótese nula de que as variáveis são independentes. Em outras palavras, temos que aceitar a hipótese nula de que o desempenho na prova de estatística independe do sexo do aluno. O CROSSTABS também calcula diretamente o nível de significância ("SIGNIFICANCE") do valor do qui-quadrado. Quando este valor fosse superior ao nível de significância escolhido (0,05) em nosso caso), teríamos que rejeitar a hipótese nula. Com isso, não seria preciso procurar o valor crítico numa tabela.

É interessante não somente saber se as variáveis são associadas ou não, mas também conhecer o *grau* de associação entre elas. O V de Cramer é uma das medidas deste grau de associação calculadas por SPSS, variando entre 0 e + 1 em função deste grau. Em nosso exemplo, o V de Cramer de 0,18 mostra um grau de associação que é relativamente baixo.

Existem várias opções para controle da inclusão ou exclusão dos MISSING VALUES nas listagens com o cartão OPTIONS (ver SPSS, pp. 241-242). Este cartão não é necessário quando o usuário quer que estes valores sejam excluídos das tabelas.

As limitações mais importantes são

(a) em relação ao *modo General*:
   (1) o número máximo de variáveis é 200,
   (2) o número máximo de valores de uma variável é 250,
   (3) um máximo de 20 "tabelas" pode ser pedido com um cartão TABLES, considerando-se a definição de "tabelas" como o conjunto de cruzamentos pedido entre duas barras,
   (4) o número máximo de dimensões é 10;

(b) em relação ao *modo Integer*
   (1) o número máximo de variáveis no cartão VARIABLES é 100,
   (2) o número máximo de variáveis em um cartão TABLES é 100,
   (3) o número máximo de valores de uma variável é 200,
   (4) o número máximo de dimensões é 8 ou 6.

6.4 BREAKDOWN — Decomposição de variáveis contínuas por uma variável categoria (SPSS, p. 249-264)

Às vezes o pesquisador quer calcular as médias e desvios-padrão para diferentes grupos dentro da mostra. Poderíamos querer, por exemplo, calcular a nota média geral na prova de estatística de todos os alunos, e ainda proceder ao cálculo das médias de diferentes subgrupos, como os de turma e sexo (ver figura 6.3). O subprograma BREAKDOWN permite este tipo de análise de uma variável contínua por uma ou mais variáveis categoriais.

AS MÉDIAS DA NOTA DE ESTATÍSTICA, SEGUNDO TURMA E SEXO DO ALUNO

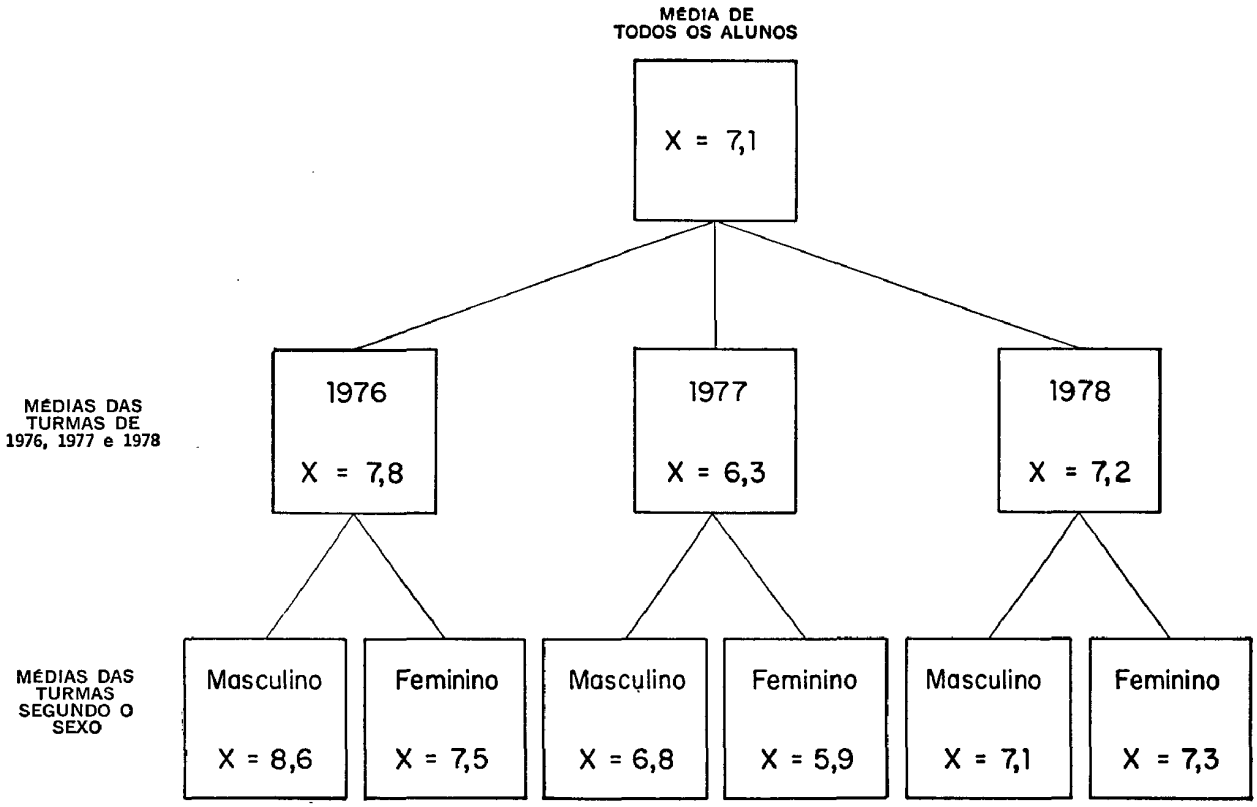


Fig. 6.3

Como o CROSSTABS, o subprograma BREAKDOWN apresenta dois modos de operar: o GENERAL e o INTEGER. Ver FREQUENCIES (6.2) para uma discussão das vantagens e desvantagens de cada modo.

1	16			
BREAKDOWN	TABLES =	variável ou lista de var.	BY	variável ou lista de var. BY ...

Poderíamos gerar os resultados para a figura 6.3 (a nota média na prova de estatística segundo turma e sexo) com o seguinte cartão:

1	16
BREAKDOWN	TABLES = EST BY TUR BY SEXO

Ou as médias e desvios-padrão para cada grupo de sexo, de turma e de área:

```
1                    16
BREAKDOWN            TABLES = EST BY SEXO BY TUR BY
                       AREA
```

Poderíamos também calcular a nota média nas provas de redação (RED), inglês (ING) e estatística (EST) segundo grupo de sexo e idade, continuando no mesmo cartão:

```
1                    16
BREAKDOWN            TABLES = EST BY SEXO BY TUR, RED,
                     EST BY SEXO BY IDADD
```

No modo de operar INTEGER deve-se incluir um cartão VARIABLES (ver formato geral do CROSSTABS). Os cartões de controle para o exemplo acima seriam:

```
1                    16
BREAKDOWN            VARIABLES = IDADE (LOWEST, HIGHEST)
                     SEXO (1,2) TUR (76,78)
                     RED, ING, EST (1,10)
                     TABLES = IDADE BY SEXO BY TUR/RED
                     ING EST BY SEXO
STATISTICS           1
OPTIONS              4
```

Cumpre salientar o uso da convenção para valores mínimos e máximos (LOWEST, HIGHEST) e as vírgulas para definir os mesmos limites para três variáveis.

Além de várias opções para inclusão dos MISSING VALUES existe uma especificamente para geração dos resultados na forma de uma árvore, como na figura 6.3. O exemplo 6.4 mostra o formato normal. Sem OPTIONS 1 a 3, ou sem o cartão OPTIONS, os MISSING VALUES são excluídos (SPSS, p. 257).

As limitações mais relevantes são (SPSS, p. 261-262).

(a) em relação ao *Modo General*:

- o número total de variáveis é 200,

- o número total de tabelas que pode ser pedido é 250,
- o número máximo de dimensões é 6, incluindo uma variável continua (dependente) e cinco variáveis categoriais (independentes);

(b) em relação ao *Modo Integer*:

- o número máximo de variáveis nos cartões TABLES é 100,
- o número máximo de tabelas pedidas é 100,
- o número máximo de dimensões é 6,
- o número máximo de "tabelas" pedidas é 30, designando-se por "tabelas" as variáveis assinaladas entre duas barras (/).

As *estatísticas* (ver SPSS, p. 257-261) além das imprimidas automaticamente (as médias, desvios-padrão, variância e número de casos) são:

1. Análise de variância com uma variável categorial (*Oneway analysis of variance*)

2. Teste de linearidade — teste de uma tendência linear (*test of linearity*).

No caso da análise de variância com uma variável categorial (ver Hoel, p. 292-296; Rodrigues, p. 136-145; Blalock, p. 317-333), pode-se testar a hipótese nula de que as médias de subpopulação definidas pelas categorias de uma variável de nível de mensuração nominal são significativamente diferentes. Ou seja, testar a hipótese nula de que:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_n$$

As suposições são as seguintes: a amostragem é aleatória, as subpopulações têm uma distribuição normal e as variâncias das subpopulações são iguais (Blalock, p. 324). Por exemplo, testaremos aqui a hipótese de que as notas de estatística são estatisticamente diferentes nas três turmas (76, 77, 78). O exemplo 6.4 apresenta o quadro de análise de variância gerado pelo número 1 (um) do cartão STATISTICS do procedimento BREAKDOWN, e a tabela 6.4 fornece algumas informações sobre a interpretação destes resultados.

TABELA 6.4

A INTERPRETAÇÃO DO QUADRO DE ANÁLISE DE VARIÂNCIA

FONTE DA VARIÂNCIA	SOMA DOS QUADRADOS	GRAUS DE LIBERDADE[1]	MÉDIA DOS QUADRADOS	F[2]	SIGNIFICÂNCIA
Entre Categorias		k — I	ENTRE		
Dentro de Categorias		N — k	DENTRO		
Total		N — I	TOTAL		

ETA SQUARED = ETA^2 = .NTRE/TOTAL.

1 j = número de categorias
N = número de observações

2 F = ENTRE/DENTRO

Para testar esta hipótese, calculamos o valor de F, a relação entre a variância entre as categorias e a que se verifica dentro delas (ver tabela 6.4). O valor crítico do F com um nível de significância de 0,05 para nosso caso com 2 graus de liberdade no numerador e 87 no denominador é 3,15. Uma vez que este valor crítico é menor do que o valor do F calculado (isto é, 8,356), podemos rejeitar a hipótese nula. Ou seja, concluímos que existem diferenças estatisticamente significativas entre as médias das três turmas na prova de estatística. O nível de significância calculado pelo SPSS é de 0,0005, ou seja, muito mais significativo do que o nível utilizado em nosso teste (0,05).

A estatística Eta^2 ("ETA SQUARED") demonstra a proporção da variância total que é "explicada" pela variável independente (ver Blalock, p. 354-353 e 410-413). Em nosso exemplo a variável independente TUR "explica" 40,12% da variância da variável dependente EST.

6.5 T-TEST — Teste de t (SPSS, p. 267-275)

O teste de t é comumente utilizado no teste de hipóteses sobre a diferença entre as médias de duas populações. Estas médias podem ser de: (1) duas amostras independentes, ou (2) duas amostras do tipo emparelhado (*paired*), como nos casos do grupo experimental e do grupo de controle, ou de um grupo com observações tomadas antes e

EXEMPLO 6.4

BREAKDOWN: CÁLCULO DA MÉDIA, DESVIO-PADRÃO E VARIÂNCIA DA VARIÁVEL EST, SEGUNDO ANO DE ENTRADA (TUR) E SEXO; TABELA DE ANÁLISE DE VARIÂNCIA DE EST COM TUR

DESCRIPTION OF SUBPOPULATIONS

CRITERION VARIABLE EST NOTA DE ESTATISTICA
BROKEN DOWN BY TUR ANO DE ENTRADA
BY SEXO SEXO DO ALUNO

VARIABLE	CODE	VALUE LABEL	SUM	MEAN	STD DEV	VARIANCE	N
FOR ENTIRE POPULATION			639.0000	7.1000	1.5726	2.4730	(90)
TUR	76.		235.0000	7.8333	1.6418	2.6954	(30)
SEXO	1.	MASCULINO	77.0000	8.5556	1.7401	3.0278	(9)
SEXO	2.	FEMININO	158.0000	7.5238	1.5368	2.3619	(21)
TUR	77.		189.0000	6.3000	1.5347	2.3552	(30)
SEXO	1.	MASCULINO	82.0000	6.8333	1.4668	2.1515	(12)
SEXO	2.	FEMININO	107.0000	5.9444	1.5136	2.2908	(18)
TUR	78.		215.0000	7.1667	1.1472	1.3161	(30)
SEXO	1.	MASCULINO	106.0000	7.0667	1.0998	1.2095	(15)
SEXO	2.	FEMININO	109.0000	7.2667	1.2228	1.4952	(15)

TOTAL CASES = 90

CRITERION VARIABLE EST

ANALYSIS OF VARIANCE

VARIABLE	CODE	VALUE LABEL	SUM	MEAN	STD DEV	SUM OF SQ	N
TUR	76.		235.0000	7.8333	1.6418	78.1667	(30)
TUR	77.		189.0000	6.3000	1.5347	68.3000	(30)
TUR	78.		215.0000	7.1667	1.1472	38.1667	(30)
	WITHIN GROUPS TOTAL		639.0000	7.1000	1.4568	184.6333	(90)

ANALYSIS OF VARIANCE

SOURCE	SUM OF SQUARES	D.F.	MEAN SQUARE	F	SIG.
BETWEEN GROUPS	35.467	2	17.733	8.356	0.0005
WITHIN GROUPS	184.633	87	2.122		

ETA = 0.4014 ETA SQUARED = 0.1611

depois da experiência (ver Rodrigues, p. 106-118 e Hoel, p. 224-228). Nesta seção analisaremos somente o primeiro tipo: o de duas amostras independentes.

A hipótese nula é que as médias de duas populações são iguais, ou seja:

$$H_0: \mu_1 = \mu_2$$

onde μ_1 e μ_2 são as médias de duas populações. Queremos testar essa hipótese nula com os dados levantados em duas amostras destas populações. Como exemplo, aqui vamos testar a hipótese nula de que as médias na prova de Estatística são iguais para os homens e as mulheres.

Há quatro etapas no teste desta hipótese: (1) formulação da hipótese nula, como foi feita acima; (2) escolha do nível de significância, ou seja, a probabilidade de se cometer um erro do tipo I (rejeitar uma hipótese quando ela deve ser aceita); (3) cálculo do valor do t com base nas amostras das duas populações e (4) o teste da hipótese nula propriamente dito.

Existem duas maneiras de se estimar o valor de t: (1) uma para as populações que têm a mesma variância ($\sigma^2 = \sigma^2$), ou seja, variância comum ou *Pooled Variance Estimate* (2), a outra para populações que têm variâncias diferentes ($\sigma^2 \neq \sigma^2$).

(*Separate Variance Estimate*). Por isso, antes de se calcular o valor de t, é preciso determinar se existe ou não uma diferença significativa entre as variâncias das duas populações. Para tanto, utiliza-se o teste de F para testar a hipótese nula de que as variâncias das duas populações são iguais, ou seja:

$$H_0: \sigma^2 = \sigma^2$$

O F é calculado com base no emprego dos dados das duas amostras, através da equação seguinte:

$$F = \frac{\text{variância } (s^2) \text{ maior dos dois grupos}}{\text{variância } (s^2) \text{ menor dos dois grupos}}$$

Se o valor do F for maior do que o valor crítico com o nível de significância escolhido (0,05), a hipótese nula de que as variâncias das duas subpopulações são iguais é rejeitada. Se rejeitamos esta hipótese nula, a estimativa de t para populações com variâncias diferentes (*Separate Variance Estimate*) deve ser utilizada. Por outro lado, no caso de se

EXEMPLO 6.5

T-TEST: TESTE DA DIFERENÇA ENTRE AS MÉDIAS DOS ALUNOS, SEGUNDO O SEXO

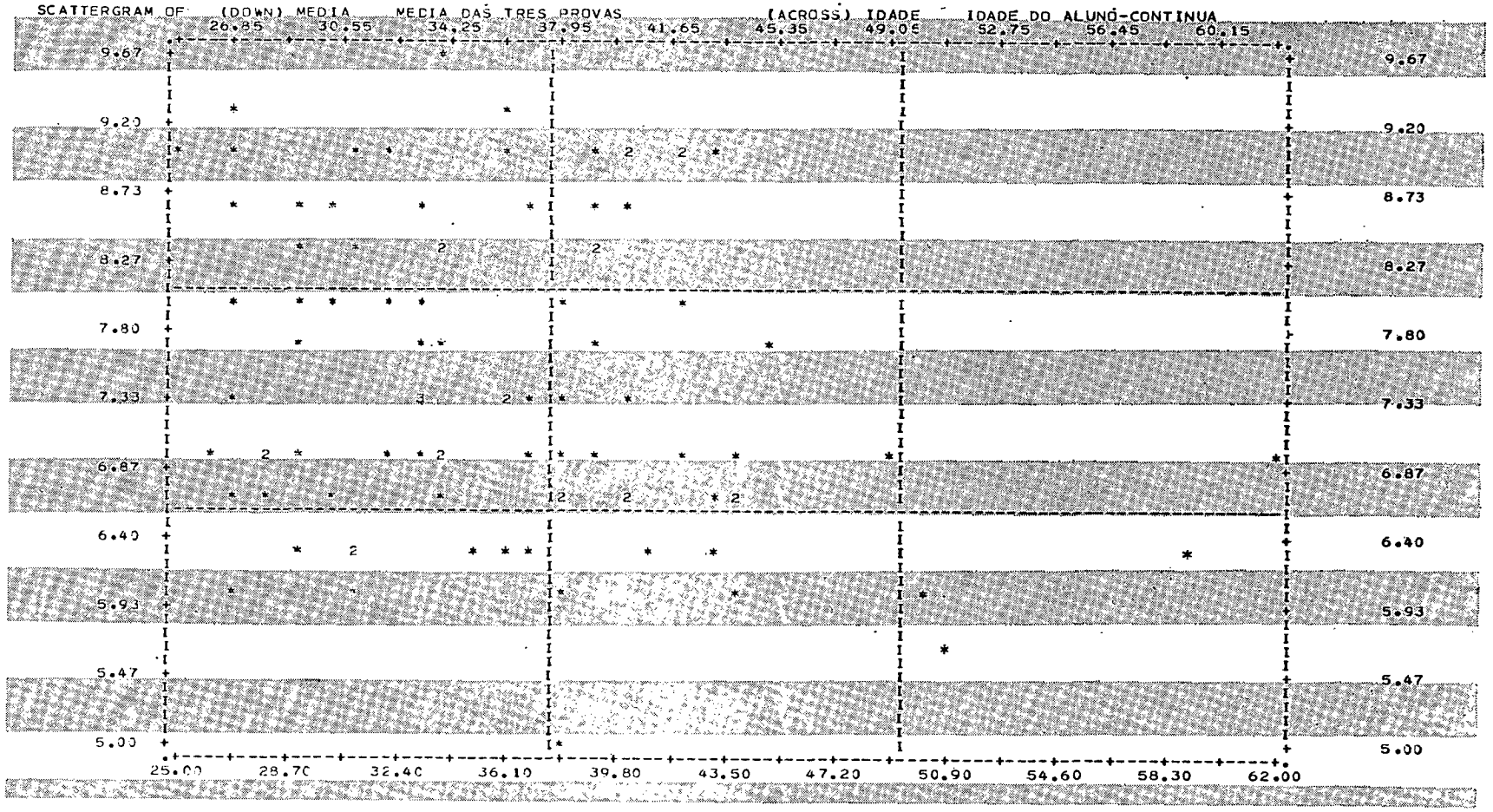
T - T E S T

GROUP 1 - SEXO EQ 1.
GROUP 2 - SEXO EQ 2.

VARIABLE		NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR	F VALUE	2-TAIL PROB.	POOLED VARIANCE ESTIMATE T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.	SEPARATE VARIANCE ESTIMATE T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
RED	NOTA DE REDACAO												
	GROUP 1	36	7.0833	1.592	0.265	1.29	0.393	-0.79	88	0.435	-0.77	68.41	0.447
	GROUP 2	54	7.3333	1.401	0.191								
ING	NOTA DE INGLES												
	GROUP 1	36	8.1944	1.670	0.278	1.13	0.686	0.24	88	0.811	0.24	72.04	0.813
	GROUP 2	54	8.1111	1.574	0.214								
EST	NOTA DE ESTATISTICA												
	GROUP 1	36	7.3611	1.533	0.256	1.07	0.837	1.29	88	0.200	1.30	76.97	0.197
	GROUP 2	54	6.9259	1.588	0.216								

EXEMPLO 6.6

SCATTERGRAM: DIAGRAMA DE DISPERSÃO DAS VARIÁVEIS MÉDIA E IDADE E OS RESULTADOS DE UMA REGRESSÃO SIMPLES



STATISTICS..

CORRELATION (R) =	-0.21669	R SQUARED =	0.04783	SIGNIFICANCE =	0.01919
STD ERR OF EST =	1.02181	INTERCEPT (A) =	8.67550	SLOPE (B) =	-0.03285
PLOTTED VALUES =	90	EXCLUDED VALUES =	0	MISSING VALUES =	0

'*********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

aceitar a hipótese nula de que as variâncias são iguais, a estimativa de t para variância comum (*Pooled Variance Estimate*) deve ser escolhida.

Em nosso exemplo (exemplo 6.5), F da variável RED é igual a 1,29, o que corresponderia a um nível de significância de 0,393, isto é, uma probabilidade de se cometer um erro do tipo I 39 vezes em 100 tentativas. Obviamente, não podemos rejeitar a hipótese nula a um nível de significância de 0,05, e, por isso, precisamos utilizar a estimativa de t com variância comum. No caso de RED, o valor de t, utilizando-se este *Pooled Variance Estimate*, é — 0,79. Com o nível de significância calculado por SPSS de 0,435, não podemos *rejeitar* a hipótese nula de que as médias das duas populações (homens e mulheres) são iguais, visto que o nível de significância escolhido aqui era de 0,05. Em outras palavras, temos que concluir que as médias das duas populações são iguais.

O cartão de controle que gerou os resultados do exemplo 6.5 é:

```
1                    16
T-TEST               GROUPS = SEXO/VARIABLES = RED,
                       ING, EST
```

O formato geral para este cartão de controle é

```
1                    16
T-TEST               GROUPS = especificação  VARIABLES =
                              da variável    variável em que
                              definindo os   as médias serão
                              dos grupos     calculadas
```

6.6 SCATTERGRAM e PEARSON CORR — Diagrama de Dispersão, Regressão Simples e Coeficiente de Correlação PEARSON (SPSS, p. 276-288, p. 293-300)

O diagrama de dispersão permite a análise visual da relação entre duas variáveis do nível de mensuração intervalar (Hoel, p. 241-271). O exemplo 6.6 demonstra a relação entre a média das três provas (MEDIA) e a idade do aluno (IDADE), sendo ambas as variáveis criadas com cartões de COMPUTE com base nos dados originais. O cartão de controle que gerou este exemplo foi:

```
1                    16
SCATTERGRAM          MEDIA WITH IDADE
STATISTICS           ALL
```

Também pode-se pedir mais que um diagrama de dispersão com um cartão:

1	16
SCATTERGRAM	MEDIA, EST, ING, WITH IDADE, RED

o que geraria 6 diagramas (MEDIA X IDADE, MEDIA X RED, EST X X IDADE, etc.).

Com o cartão STATISTICS pode-se gerar as estatísticas de uma regressão linear simples:

1. Coeficiente de Correlação Pearson (r) — CORRELATION (R)
2. R quadrado — R SQUARED
3. Significância de r — SIGNIFICANCE
4. Erro-Padrão da Estimativa — STD ERR OF EST
5. Intercessão com o eixo vertical — INTERCEPT
6. Inclinação — SLOPE

Poder-se-ia, então, expressar a relação ajustada entre MEDIA e IDADE pela equação linear:

MEDIA = 8,676 — 0,033 IDADE

o R^2 mostra que a variável independente IDADE explica 0,04783% da variância da variável dependente MEDIA. O teste de t unilateral mostra um nível de significância do r de 0,01919, o que comprova que a probabilidade de se cometer um erro de tipo I seria 0,019 no caso de se rejeitar a hipótese nula ($r = 0$).

Também pode-se calcular somente o coeficiente de correlação Pearson e o teste de t de sua significância com o procedimento PEARSON CORR. O Exemplo 6.7 foi gerado com o seguinte cartão:

1	16
PEARSON	RED, EST, ING, IDADE

Na listagem, a significância aparece depois da letra S. Por exemplo, o coeficiente de correlação entre as variáveis EST e IDADE é — 0,2871. Poderíamos rejeitar a hipótese nula de que não existe correlação nenhuma entre estas duas variáveis ($r = 0$) com um nível de significância de 0,003.

EXEMPLO 6.7

PEARSON CORR: COEFICIENTES DE CORRELAÇÃO PEARSON ENTRE AS VARIÁVEIS ING, EST, RED, IDADE E MÉDIA

	ING	EST	RED	IDADE	MEDIA
ING	1.0000 (0) S=0.001	0.2970 (90) S=0.002	0.0046 (90) S=0.483	-0.1284 (90) S=0.114	0.6653 (90) S=0.001
EST	0.2970 (90) S=0.002	1.0000 (0) S=0.001	0.2172 (90) S=0.020	-0.2825 (90) S=0.003	0.7587 (90) S=0.001
RED	0.0046 (90) S=0.483	0.2172 (90) S=0.020	1.0000 (0) S=0.001	-0.0222 (90) S=0.418	0.5845 (90) S=0.001
IDADE	-0.1284 (90) S=0.114	-0.2825 (90) S=0.003	-0.0222 (90) S=0.418	1.0000 (0) S=0.001	-0.2187 (90) S=0.019
MEDIA	0.6653 (90) S=0.001	0.7587 (90) S=0.001	0.5845 (90) S=0.001	-0.2187 (90) S=0.019	1.0000 (0) S=0.001

6.7 NONPAR CORR — Correlação Não-paramétrica Spearman ou Kendall (SPSS, p. 288-292)

Os coeficientes de correlação Spearman e Kendall são estatísticas não-paramétricas que mostram o grau de associação entre duas variáveis de nível de mensuração ordinal (Blalock pp. 415-428 e Spiegel, pp. 429-431). Por exemplo, calcularemos estes coeficientes entre as notas das três provas e aqueles entre elas e a idade. Os coeficientes de correlação Spearman e Kendall, do mesmo modo que o de Pearson, variam entre — 1,0 e + 1,0 (ver tabela 6.5). Um coeficiente de + 1,0 significa que as duas variáveis têm ordenações exatamente iguais, e um coeficiente — 1,0, que as duas variáveis têm ordenações exatamente inversas. E um coeficiente de valor zero significa que não existe nenhuma relação ordinal entre as ordenações das duas variáveis.

TABELA 6.5

OS COEFICIENTES DE CORRELAÇÃO SPEARMAN E KENDALL COM DIFERENTES ORDENAÇÕES DA SEGUNDA VARIÁVEL

ORDENAÇÃO DA PRIMEIRA VARIÁVEL	ORDENAÇÃO DA SEGUNDA VARIÁVEL	
1	1	5
2	2	4
3	3	3
4	4	2
5	5	1
Coeficientes de correlação	+ 1,0	— 1,0

O cartão seguinte gera os resultados do exemplo 6.7:

```
1                    16
NONPAR CORR          ING, RED, EST, IDADE
OPTIONS              6
```

Note-se que o procedimento NONPAR CORR não usa o cartão STATISTICS. Para calcular só o coeficiente de Kendall recorre-se à opção 5; para produzir simultaneamente os coeficientes de Kendall e Spearman,

EXEMPLO 6.8

NONPAR CORR: COEFICIENTE DE CORRELAÇÃO SPEARMAN E KENDALL PARA AS VARIÁVEIS ING, RED, EST E IDADE

- - - - - - - - - - - - - - - SPEARMAN CORRELATION COEFFICIENTS - - - - - - - - - - - - - -

VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR	
ING WITH RED	0.0512 N(90) SIG .316	ING WITH EST	0.3141 N(90) SIG .001	ING WITH IDADE	-0.0206 N(90) SIG .424	RED WITH EST	0.2539 N(90) SIG .008	RED WITH IDADE	-0.0102 N(90) SIG .462	EST WITH IDADE	-0.2480 N(90) SIG .009

- - - - - - - - - - - - - - - KENDALL CORRELATION COEFFICIENTS - - - - - - - - - - - - - -

VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR		VARIABLE PAIR	
ING WITH RED	0.0372 N(90) SIG .331	ING WITH EST	0.2488 N(90) SIG .002	ING WITH IDADE	-0.0172 N(90) SIG .415	RED WITH EST	0.1975 N(90) SIG .009	RED WITH IDADE	-0.0151 N(90) SIG .425	EST WITH IDADE	-0.1784 N(90) SIG .012

à opção 6. Sem o cartão OPTIONS, somente os coeficientes Spearman serão gerados. Em geral, o coeficiente Kendall é considerado mais fidedigno quando o número de empates entre as ordenações for maior. A listagem também mostra a significância do coeficiente, calculada com um teste de t unilateral.

7. A ORDEM DOS CARTÕES DE CONTROLE DO SPSS NO "DECK" E A LISTAGEM DOS CARTÕES EMPREGADOS NO EXEMPLO

A tabela 7.1 apresenta a ordem dos cartões opcionais e dos obrigatórios no "deck" de SPSS. E o exemplo 7.1 mostra o deck utilizado para fazer muitas das operações discutidas anteriormente. O usuário deve consultar seu núcleo de computação eletrônica para saber o JCL

TABELA 7.1

ORDEM DOS CARTÕES DE CONTROLE DOS SPSS

TIPO OPCIONAL OU NÃO	CARTÃO DE CONTROLE	COMENTÁRIOS
A. Opcional	RUN NAME	Cartões de definição de dados.
Obrigatório	VARIABLE LIST	
Obrigatório	INPUT MEDIUM	
Obrigatório	N OF CASES	
Obrigatório		
Obrigatório	INPUT FORMAT	
Opcional	MISSING VALUES	Estes cartões podem aparecer em qualquer ordem, inclusive depois de cartões do tipo B.
Opcional	VAR LABELS	
Opcional	VALUE LABELS	
B. Opcional	Cartões de criação e transformação de variáveis (por exemplo, COMPUTE, SELECT IF, IF, RECODE, etc.).	
C. Opcional	LIST CASES	Listagem dos dados
D. Obrigatório	Cartão de Operação	CROSSTABS, BREAKDOWN, etc.
	OPTIONS	Cartões de Definição da Primeira Operação
	STATISTICS	
E. Obrigatório	READ INPUT DATA	
F.	Quando são utilizados cartões como o INPUT MEDIUM, os cartões de dados devem ser colocados aqui.	
G.	Outros cartões de Operações podem ser aqui utilizados. Note-se que eles vêm depois dos cartões de dados.	
	Cartões de Operação	
	OPTIONS	
	STATISTICS	
	Cartão de Operação	
	OPTIONS etc.	
H.	FINISH	

EXEMPLO 7.1

A LISTAGEM DOS CARTÕES DE CONTROLE DO EXEMPLO AQUI UTILIZADOS

Cartões de JCL (Job Control Language) *

```
RUN NAME        DADOS SOBRE AS TURMAS DE 76 77 78
VARIABLE LIST   MAT,SEXO,DIA,MES,ANO,AREA,TUR,ING,EST,RED
INPUT MEDIUM    CARD
N OF CASES      90
INPUT FORMAT    FIXED(F6.0,F2.0,3F3.0,F2.0,4F3.0)
COMPUTE         IDADE=79-ANO
COMPUTE         IDCAT=IDADE
RECODE          IDCAT (LOWEST THRU 24=1) (25 THRU 29=2) (30 THRU 34=3),
                (35 THRU 39=4) (40 THRU HIGHEST=5)
COMPUTE         MEDIA=(EST+ING+RED)/3.0
VAR LABELS      MAT,NUMERO DA MATRICULA/
                SEXO,SEXO DO ALUNO/
                DIA,DIA DE NASCIMENTO/
                MES,MES DE NASCIMENTO/
                ANO,ANO DE NASCIMENTO/
                IDADE,IDADE DO ALUNO-CONTINUA/
                IDCAT,IDADE DO ALUNO-CATEGORIAL/
                AREA,AREA DE CONCEN/
                TUR,ANO DE ENTRADA/
                MEDIA,MEDIA DAS TRES PROVAS/
                ING,NOTA DE INGLES/
                EST,NOTA DE ESTATISTICA/
                RED,NOTA DE REDACAO
VALUE LABELS    SEXO (1)MASCULINO (2)FEMININO/
                TUR (1)TURMA DE 76 (2)TURMA DE 77 (3)TURMA DE 78/
                AREA (1)PLANEJ (2)MET E TEC (3)ACONS/
                IDCAT (1)MENOS DE 25 ANOS (2)DE 25 A 30 (3)DE 30 A 35
                (4)DE 35 A 40 (5)DE 40 E MAIS
                /EST,ING,RED (6) ATE 6 (9) 9 +
LIST CASES      CASES = 90/ VARIABLES = SEXO,ANO,IDADE
CONDESCRIPTIVE  RED,EST,ING
STATISTICS      ALL
STATISTICS      ALL
READ INPUT DATA
16041.1.14.04.43.2.76.10.10.08.
16042.1.21.10.40.1.76.10.09.07.
16043.1.21.06.52.2.76.10.10.08.
16045.1.08.04.37.3.76.08.10.09.
        .
        .
        .
16061.1.16.10.48.2.78.08.06.05.
16062.1.05.08.50.3.78.06.06.07.
FREQUENCIES     GENERAL=SEXO,AREA,TUR,RED,ING,EST
STATISTICS      ALL
OPTIONS         8
*RECODE         EST,ING,RED(4,5=6)(10=9)
CROSSTABS       TABLES=SEXO BY RED,ING,EST
STATISTICS      1,2
CROSSTABS       VARIABLES=SEXO(1,2) IDCAT(1,5) TUR(76,78)/
                TABLES=SEXO,IDCAT BY TUR/SEXO BY IDCAT BY TUR
BREAKDOWN       TABLES=EST BY SEXO BY TUR
STATISTICS      ALL
BREAKDOWN       TABLES=RED,EST,ING BY SEXO,TUR,APEA,IDCAT /
                EST BY TUR/
                EST BY TUR BY SEXO/
                EST BY SEXO BY TUR
STATISTICS      1
BREAKDOWN       VARIABLES=IDADE(LOWEST,HIGHEST) SEXO(1,2) TUR(76,78)/
                TABLES=IDADE BY SEXO BY TUR
T-TEST          GROUPS=SEXO / VARIABLES=RED, ING, EST
SCATTERGRAM     MEDIA WITH IDADE
STATISTICS      ALL
SCATTERGRAM     ING,RED,EST WITH IDADE.
PEARSON CORR    ING,EST,RED,IDADE,MEDIA
STATISTICS      ALL
NONPAR CORR     ING,RED,EST,IDADE
OPTIONS         4,6
FINISH
```

* (Job Control Language) necessário para rodar o SPSS. Os cartões de JCL, que aparecem antes do deck dos cartões de controle do SPSS, apresentam várias informações necessárias para contabilidade e controle do sistema de computação, como também a de que o usuário quer utilizar o SPSS, com ou sem fitas ou discos magnéticos para a entrada dos dados a serem processados.

8. UMA NOTA FINAL

A utilização das informações aqui apresentadas depende muito do nível de sofisticação estatística dos alunos e de suas preocupações. Para um nível introdutório normalmente apresentamos as técnicas através da utilização de um exemplo bem simples, como foi feito aqui. Os alunos rodam seus programas e interpretam os resultados. Via de regra, pedimos que eles gerem várias cópias destes resultados para serem utilizadas como apostilas durante as aulas (normalmente um parâmetro no cartão JOB determina o número de cópias geradas).

A metodologia pode ser outra em se tratando de uma turma de alunos que já esteja trabalhando com os dados de sua tese ou de uma pesquisa. Neste último caso esclarecemos somente os aspectos relevantes para cada pesquisa, deixando o aluno aprofundar seu conhecimento de SPSS com as informações aqui apresentadas e as do manual de SPSS.

Visto que cada professor e cada turma são diferentes, tentei apresentar uma seleção das técnicas mais utilizadas que poderiam constituir a fundação de um conhecimento maior da análise estatística nas ciências com o SPSS. Espero que estas informações venham a servir para facilitar a aplicação deste método, reduzindo-se dessa maneira, um pouco, o hiato entre a teoria e a prática de métodos quantitativos empregados nas ciências sociais.

BIBLIOGRAFIA

BLALOCK, H. M. *Social Statistics* (Mexico: McGraw-Hill Kogakusha, 1972).

CASTRO, Claudio de Moura. *A Prática de Pesquisa* (Rio de Janeiro: McGraw-Hill, 1977).

CALDEIRA, A. M. Salgueira e Ferreira, M. L. de Brito M. G. *Estatística: Instrução Programada.* Volumes 1 e 2 (Rio de Janeiro: CONQUISTA, 1975).

HOEL, P. G. *Estatística Elementar* (São Paulo: Atlas, 1977).

KIRK, R. E. *Statistical Issues: A Reader for the Behavioral Sciences* (Monterey, California: Brooks/Cole, 1972).

KLECKA, W. R. *et alii. SPSS Primer* (New York: McGraw-Hill, 1975).

NIE, N. H. *et alii. SPSS: Statistics Package for the Social Sciences* (New York: McGraw-Hill, 1975).

RIO DATA CENTRO, *Texto Básico para Palestra de Introdução ao SPSS*, PUC, Rio de Janeiro (este texto está arquivado no sistema de documentação do Rio Data Centro (LISTADOC). O usuário pode pedir uma cópia dele com o cartão JOB e o seguinte cartão: EXEC LISTADOC, NOME = SPSS2).

SIEGEL, S. *Estatística Não Paramétrica* (São Paulo: McGraw-Hill, 1977).

SPIEGEL, M. R. *Estatística* (Rio de Janeiro: McGraw-Hill, 1977).

RODRIGUES, A. *A Pesquisa Experimental em Psicologia e Educação* (Petrópolis: Vozes, 1975).

Comunicações

EMPREGO DA FUNÇÃO FATORIAL PARA O CÁLCULO DOS MOMENTOS DE UMA DISTRIBUIÇÃO DE FREQÜÊNCIA

Hélio Ventura da Cruz *

SUMÁRIO

1. INTRODUÇÃO

O presente estudo apresenta, para o cálculo dos momentos ordinários (e centrais) de uma distribuição de freqüência, com intervalos de classes de amplitude constante, uma abordagem pouco empregada nos compêndios de estatística descritiva, que é a utilização dos momentos fatoriais. Para o cálculo desses momentos são utilizadas somente somas de freqüências acumuladas, evitando-se, desta forma, o cálculo dos produtos das potências dos pontos médios das classes pelas respectivas freqüências.

* Professor da Escola Nacional de Ciências Estatísticas — ENCE/IBGE.

Além disso, em relação ao processo usual a ordem de grandeza dos números utilizados nos cálculos dos momentos fatoriais é tanto menor quanto maior for o número de intervalos de classe.

Vejamos, a seguir, alguns conceitos teóricos em que se baseia o processo de cálculo.

2. CONCEITOS BÁSICOS

2.1 Função Fatorial

Dá-se o nome de Função Fatorial de grau k, e se representa por $(x)_k$, ao seguinte polinômio incompleto de grau k.

$$(x)_k = x\ (x-1)\ .\ (x-2)\ \ldots\ (x-k+1) \tag{1}$$

Uma vez definida a Função Fatorial, é sempre possível obter-se qualquer potência inteira k de uma variável x como combinação linear das funções fatoriais de graus k, $k-1$ ..., 2 e 1.

De fato, é fácil ver que

$$\begin{aligned} x &= (x)_1 \\ x^2 &= (x)_1 + (x)_2 \\ x^3 &= (x)_1 + 3\ (x)_2 + (x)_3 \\ x^4 &= (x)_1 + 7\ (x)_2 + 6\ (x)_3 + (x)_4 \end{aligned} \tag{2}$$

De um modo geral, podemos escrever

$$x^k = S_k^1\ (x)_1 + S_k^2\ (x)_2\ \ldots\ + S_k^k\ (x)_k \tag{3}$$

onde as constantes S_k^i são denominadas números de Stirling de 2.ª espécie e que são obtidos pela seguinte relação de recorrência.

$$S_k^i = S_{k-1}^{i-1} + i\ S_{k-1}^i$$

e

$$S_k^k = S_k^1 = 1 \tag{4}$$

NÚMEROS DE STIRLING DE 2.ª ESPÉCIE

k \ i	1	2	3	4	5
1	1	1			
2	1	1			
3	1	3	1		
4	1	7	6	1	
5	1	15	25	10	1

2.2 Momento Fatorial

Denomina-se momento fatorial de ordem k, e se representa por $\alpha_{(k)}$, a expectância de $(x)_k$, isto é:

$$\alpha_{(k)} = E\{(x)_k\} = \begin{cases} \Sigma\ (x)_k\ p(x) & \text{se } x \text{ é do tipo discreto} \\ \int (x)_k\ .\ f(x) & \text{se } x \text{ é do tipo contínuo} \end{cases}$$

então, quando x é do tipo discreto tem-se

$$\alpha_{(1)} = \sum_{x=1}^{n} (x)_1,\ p(x) = \sum_{x=1}^{n} x\ .\ p(x) = m \quad \text{que é média de } x$$

$$\alpha_{(2)} = \sum_{x=1}^{n} (x)_2\ .\ p(x) = \sum_{x=1}^{n} x(x\text{-}1)\ \ (p(x) = \sum_{x=1}^{n} A_x^2\ .\ p(x)$$

$$\alpha_{(3)} = \sum_{x=1}^{n} (x)_3\ .\ p(x) = \sum_{x=1}^{n} x\ .\ (x\text{-}1)\ .\ (x\text{-}2)\ .\ p(x) = \sum_{x=1}^{n} A_3^x\ p(x)$$

e, de um modo geral,

$$\alpha_{(k)} = \sum_{x=k}^{n} (x)_k\ p(x) = \sum_{x=k}^{n} A_x^k\ p(x) \qquad (5)$$

visto que $(x)_k = 0$ se $x \leqslant k$

2.3 Relação entre Momento Fatorial e Momento Ordinário

Sabe-se que o momento ordinário de ordem k, α_k, é definido como a expectância de x^k, isto é,

$$\alpha_k = E(x^k) = \begin{cases} \Sigma\ x^k\ p\ (x) & \text{se } x \text{ é discreto} \\ \int x^k\ f(x)\ dx & \text{se } x \text{ é contínuo} \end{cases}$$

tendo em vista a (3) tem-se

$$E(x^k) = E\{S_k^1 (x)_1 + S_k^2 (x)_2 + \ldots + S_k^k (x)_k\}$$

ou

$$E(x^k) = S_k^1 E(x)_1 + S_k^2 E(x)_2 + \ldots + S_k^k E(x)_k$$

e pelas definições de momento fatorial e momento ordinário vem finalmente

$$\alpha_k = S_k^1 \alpha_{(1)} + S_k^2 \alpha_{(2)} + \ldots + S_k^k \alpha_{(k)} \tag{6}$$

Fazendo-se $k = 1, 2, 3, 4$ obtem-se

$$\begin{aligned} \alpha_1 &= \alpha_{(1)} = m \text{ que é a média de } x \\ \alpha_2 &= \alpha_{(1)} + \alpha_{(2)} \\ \alpha_3 &= \alpha_{(1)} + 3\,\alpha_{(2)} + \alpha_{(3)} \\ \alpha_4 &= \alpha_{(1)} + 7\,\alpha_{(2)} + 6\,\alpha_{(3)} + \alpha_{(4)} \end{aligned} \tag{7}$$

2.4 Relação entre os Momentos Centrais e Ordinários

Sabe-se que o Momento Central de ordem k, que se representa por μ_k é definido como:

$$\mu_k = E\{x - E(x)\}^k = E(x - m)^k$$

é óbvio que:

$$\mu_0 = 1$$

$$\mu_1 = 0$$

O Momento Central de 2.ª ordem, μ_2 é denominado variância e a sua raiz quadrada de desvio-padrão, isto é

Variância

$$\sigma^2 = \mu_2 = E(x - m)^2$$

Desvio padrão

$$\sigma = \sqrt{\mu_2}$$

vejamos as relações entre μ_k e α_k,

para $k = 2, 3$ e 4 tem-se

$$\mu_2 = E(x - m)^2 = E(x^2 - 2m\,x + m^2) = E(x^2) - 2m\,E(x) + m^2$$

$$\mu_3 = E(x - m)^3 = E(x^3 - 3m\,x^2 + 3m^2x - m^3) =$$

$$= E(x^3) - 3m\,E(x^2) + 3m^2\,E(x) - m^3$$

$$\mu_4 = E(x - m)^4 = E(x^4) - 4m\,E(x^3) +$$

$$+ 6\,m^2\,E(x^2) - 4m^3\,E(x) + m^4$$

Então, pela definição de Momento Ordinário, obtém-se

$$\begin{aligned} \mu_2 &= \alpha_2 - \alpha_1^2 \\ \mu_3 &= \alpha_3 - 3\,\alpha_2\,\alpha_1 + 2\,\alpha_1^3 \\ \mu_4 &= \alpha_4 - 4\,\alpha_3\,\alpha_1 + 6\,\alpha_2\,\alpha_1^2 - 3\,\alpha_1^4 \end{aligned} \tag{8}$$

2.5 Relação entre os Momentos Centrais e Fatoriais

Substituindo-se as expressões de (7) em (8) obtém-se:

$$\begin{aligned} \mu_2 &= \alpha_{(2)} - m\,(m-1) \\ \mu_3 &= \alpha_{(3)} - 3\,(m-1)\,\alpha_{(2)} + m\,(m-1)\,(2m-1) \\ \mu_4 &= \alpha_{(4)} - 2\,\{2\,(m-1) - 1\}\,\alpha_{(3)} + \\ &\quad + \{6\,(m-1)^2 + 1\}\,\alpha_{(2)} - m\,(m-1)\,\{1 + 3m\,(m-1)\} \end{aligned} \tag{9}$$

3. DISTRIBUIÇÃO DE FREQÜÊNCIA (DF)

3.1 Elementos de uma DF

x_i — Ponto médio da classe i

h — Amplitude do intervalo de classe

n — Números de intervalos de classe

f_i — Freqüência absoluta simples de classe i

F_i — Freqüência absoluta acumulada (do tipo) $\geqslant$ de classe i

$$N = \sum_{i=1}^{n} f_i$$

3.2 Momentos de uma DF

Fazendo-se $Y_i = \dfrac{X_i - X_1 + h}{h}$ então, $Y_i = 1, 2, \ldots n$ é uma variável discreta, além disso, considera-se, como aproximação, que

$$p(y) = \frac{fi}{n}$$

e neste caso, os momentos ordinários central e fatorial de ordem k assumem as seguintes expressões:

$$\mu_k = \frac{1}{N} \sum_{i=1}^{n} (y_i - m)^k f_i$$

$$\alpha_k = \frac{1}{N} \sum_{i=1}^{n} (y_i)_k f_i \qquad (10)$$

3.3 Freqüência acumulada de ordem k de uma DF

Denomina-se freqüência acumulada de ordem k, da i ésima classe, que representa-se por $F_{i,k}$ ao seguinte somatório:

$$F_{i,k} = \sum_{i=1}^{n} F_{j,k-1} \qquad \begin{matrix} i = 1, 2, \ldots n \\ k = 1, 2, 3, \ldots \end{matrix} \qquad (11)$$

onde

$$F_{i,0} = F_i = \sum_{j=i}^{n} f_j$$

É sempre possível explicitar qualquer $F_{i,k}$ em função de F_i ou de f_i. Assim é que, fazendo-se $k = 1$ em (11) vem

$$F_{1,1} = \sum_{j=1}^{n} F_{j,0} = \sum_{j=1}^{n} F_j$$

$$F_{2,1} = \sum_{j=2}^{n} F_{j,0} = \sum_{j,2}^{n} F_j$$

$$F_{n,1} = \sum_{j=n}^{n} F_{j,0} = F_n$$

podemos então escrever

$$F_{i,1} = \sum_{j=1}^{n} A_{j-1}^{o} F_j \quad i = 1, 2, \ldots, n \qquad (12)$$

quando $k = 2$ tem-se

$$F_{1,2} = \sum_{j=1}^{n} F_{j,1} = F_{1,1} + F_{2,1} + \ldots + F_{n,1} =$$

$$= (F_1 + F_2 + \ldots + F_n) + (F_2 + F_3 + \ldots + F_n) + \ldots +$$

$$+ (F_{n-1} + F_n) + F_n = F_1 + 2\ F_2 + 3\ F_3 + \ldots + n\ F_n$$

$$F_{2,2} = \sum_{j=2}^{n} F_{j,2} = F_{2,1} + F_{3,1} + \ldots + F_{n,1} = (F_2 + F_3 + \ldots$$

$$+ F_n) + (F_3 + F_4 + \ldots + F_n) + \ldots (F_{n-1} - F_n) +$$

$$+ F_n = F_2 + 2\ F_3 + 3\ F_4 + \ldots (n\text{-}1)\ F_n$$

$$F_{3,2} = \sum_{j=3}^{n} F_{j,1} = F_{3,1} + F_{4,1} + \ldots + F_{n,1} =$$

$$= F_3 + 2\ F_4 + 3\ F_5 + \ldots (n\text{-}2)\ F_n$$

e de um modo geral tem-se

$$F_{i,2} = \sum_{j=i}^{n} A^{1}_{j-i+1}\ F_j \quad i = 1, 2, \ldots, n \tag{13}$$

Por um raciocínio análogo chega-se que:

$$F_{i,3} = \frac{1}{2!} \sum_{j=i}^{r} A^{2}_{j-i+2}\ F_j \quad i = 1, 2, \ldots n \tag{14}$$

obtendo-se finalmente a relação de ordem k

$$F_{i,k} = \frac{1}{(k-1)!} \sum_{j=i}^{n} A^{k-1}_{j-i+k-1}\ F_j \qquad \begin{matrix} i = 1, 2, \ldots n \\ k = 1, 2, 3, \ldots \end{matrix} \tag{15}$$

Para se obter $F_{i,k}$ em função das freqüências simples, basta desenvolver a (15) para valores de k e i e substituir

F_j por $\sum_{j=k}^{n} f_j$, obtendo-se:

$$F_{i,1} = \sum_{j=i}^{n} (j - i + 1)\ f_j \quad i = 1, 2, \ldots n$$

$$F_{i,2} = \frac{1}{2!} \sum_{j=i}^{n} A^{2}_{j-i+2}\ f_j \quad i = 1, 2, \ldots, n$$

$$F_{i,3} = \frac{1}{3!} \sum_{j=i}^{n} A^{3}_{j-i+3}\ f_j \quad i = 1, 2, \ldots n$$

e, de um modo geral, tem-se

$$F_{i,k} = \frac{1}{k!} \sum_{j=i}^{n} A_{j-i+k}^{k} f_j \qquad \begin{matrix} i = 1, 2, \ldots n \\ K = 1, 2, 3, \ldots \end{matrix} \tag{16}$$

4. CÁLCULOS DOS MOMENTOS DE UMA DF

4.1 Momentos Fatoriais

O cálculo dos momentos fatoriais está intimamente ligado ao dos $F_{i,k}$. De fato, fazendo-se em (16) $i = k$ obtém-se

$$F_{k,k} = \frac{1}{k!} \sum_{j=k}^{n} A_j^k f_j \quad k = 1, 2, 3, \ldots$$

ou

$$\sum_{j=1}^{n} A_j^k f_j = k! F_{k,k}^k \quad K = 1, 2, 3, \ldots$$

e pela (10)

$$\frac{1}{N} \sum_{j=k}^{n} A_j^k f_j = \alpha_{(k)}$$

e conseqüentemente

$$\alpha_{(k)} = \frac{k! F_{k,k}}{N} \tag{17}$$

que é a fórmula de cálculo dos momentos fatoriais.

4.2 Momentos Ordinários de uma DF

Os momentos ordinários podem ser também obtidos em função de $F_{i,k}$. Assim, fazendo-se em (17) $k = 1, 2, 3, 4$ e substituindo-se em (7) obtém-se

$$\alpha_1 = \frac{F_{1,1} + F_{1,1}}{N}$$

$$\alpha_2 = \frac{F_{1,1} + 2 F_{2,2}}{N} \tag{18}$$

$$\alpha_3 = \frac{F_{1,1} + 6\ F_{2,2} + 6\ F_{2,3}}{N}$$

$$\alpha_4 = \frac{F_{1,1} + 14\ F_{2,2} + 36\ F_{3,3} + 24\ F_{4,4}}{N}$$

4.3 Momentos Centrais

Para se obter os momentos centrais em função de $F_{i,k}$, basta substituir em (8) as expressões de (18) encontrando-se:

$$\mu_2 = \frac{2\ F_{2,2} - F_{1,1}\ (F_{1,1} - 1)}{N}$$

$$\mu_3 = \frac{6\ F_{3,3} - 6\left(\frac{F_{1,1}}{N} - 1\right) F_{2,2} + \left(\frac{F_{1,1}}{N} - 1\right)\left(2\ \frac{F_{1,1}}{N} - 1\right) F_{1,1}}{N} \qquad (19)$$

$$\mu_4 = \frac{24\ F_{4,4} - 12\left\{2\left(\frac{F_{1,1}}{N} - 1\right)\right\} F_{3,3} + 2\left\{6\left(\frac{F_{1,1}}{N} - 1\right)^2 + 1\right\} F_{2,2} - \left(\frac{F_{1,1}}{N} - 1\right)\left\{1 + \frac{3\ F_{1,1}}{N}\left(\frac{F_{1,1}}{N} - 1\right)\right\}\frac{F_{1,1}}{N}}{N}$$

4.4 Tabela de Cálculos

Para o cálculo dos 4 primeiros momentos fatoriais constrói-se a seguinte tabela de cálculos:

y	f_1	F_1	$F_{1,1}$	$F_{1,2}$	$F_{1,3}$	$F_{1,4}$
1	f_1	F_1	$\underline{F_{1,1}}$	$F_{1,2}$	$F_{1,3}$	$F_{1,4}$
2	f_2	F_2	$F_{2,1}$	$\underline{F_{2,2}}$	$F_{2,3}$	$F_{2,4}$
3	f_3	F_3	$F_{3,1}$	$F_{3,2}$	$\underline{F_{3,3}}$	$F_{3,4}$
4						$\underline{F_{4,4}}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
N	f_n	F_n	$F_{n,1}$	$F_{n,2}$	$F_{n,3}$	$F_{n,4}$
Σ	$N = f_1$	$F_{1,1}$	$F_{1,2}$	$F_{1,3}$	$F_{1,4}$	$F_{1,5}$

onde $F_{i,k}$ é dado pela (15)

4.5 Momentos Corrigidos

Observe-se que os momentos até então calculados dizem respeito a uma variável reduzida do tipo:

$$y_i = \frac{x_i - K}{h}$$

Então para obtermos os momentos da DF variável x_i devem-se corrigir os já encontrados para a variável y_i. Sejam então, α'_r e μ'_r, os momentos ordinários e centrais da DF, isto é, de variável

$$x = hy + K$$

então, por definição tem:

$$\alpha'_r = E(x^r)$$
$$\mu'_x = E\{X - E(x)\}^r$$

logo, para os 4 primeiros momentos ordinários tem-se

$$\mu = \alpha'_1 = E(x) = E(h\ x + K)$$

$$\alpha'_1 = E(x^2) = E(hx + K)^2 = E(h^2\ x^2 + 2hKx + K^2) \qquad (20)$$

$$\alpha'_3 = E(x^3) = E(hx + K)^3 = E(h^3\ x^3 + 3h^2\ Kx^2 + 3hK^2 + K^3)$$

$$\alpha'_4 = E(x^4) = E(hx + K)^4 = E(h^4\ x^4 + 4h^3\ Kx^3 + 6h^2\ K^2\ x^2 + 4hK^3\ x + K^4)$$

ou

$$\alpha'_1 = \mu = h\ \alpha_1 + K$$

$$\alpha'_2 = h^2\ \alpha_2 + 2h\ K\ \alpha_1 + K^2$$

$$\alpha'_3 = h^3\ \alpha_3 + 3h^2\ K\ \alpha_2 + 3h\ K^2\ \alpha_1 + K^3 \qquad (21)$$

$$\alpha'_4 = h^4\ \alpha_4 + 4h^3\ K\ \alpha_3 + 6h^2\ K^2\ \alpha_2 + 4h\ K^3\ \alpha_1 + K^4$$

relações que fornecem os momentos ordinários corrigidos em função dos momentos ordinários da variável reduzida y.

Por outro lado, aplicando-se a (2) em (20) e tendo em vista a definição de momento fatorial, obtém-se:

$$\alpha'_1 = \mu = h\ \alpha_{(1)} + K$$

$$\alpha'_2 = h^2\ \alpha_{(2)} + \{(h+K)^2 - K^2\}\ \alpha_{(1)} + K^2 \qquad (22)$$

$$\alpha'_3 = h^3\ \alpha_{(3)} + 3h^2\ (h+K)\ \alpha_{(2)} + \{(h+K)^3 - K^3\}\ \alpha_{(1)} + K^3$$

$$\alpha'_4 = h^4\ \alpha_{(4)} + 2h^3\ \{2\ (h+K) + h\}\ \alpha_{(3)} + h^2$$
$$\{6\ (h+K)^2 + h^2\}\ \alpha_{(2)} + \{(h+K)^4 - K^4\}\ \alpha_{(1)} + K^4$$

relações estas que fornecem os momentos corrigidos em função dos momentos fatoriais de x.

De modo análogo, para os momentos centrais tem-se:

$$\mu'_r = E\ (x - n)^r = E\ (hx + k - hm - k)^r = h^r\ E(x - m)^r$$

logo

$$\mu'_r = h^r\ \mu_r \qquad (23)$$

que é a relação que fornece os momentos centrais corrigidos em função dos momentos centrais da variável transformada x.

4.6 Exemplo

A título de aplicação, com cálculos os momentos da seguinte DF.

PONTO MÉDIO DA CLASSE x_i	FREQÜÊNCIA
125	108
135	124
145	152
155	170
165	158
175	139
185	128
195	21
Σ	1000

De acordo com o item 4.4 tem-se o quadro de calcular auxiliares.

TABELA 1

$Y_i = \frac{x_i - 115}{10}$	f_i	F_1	$F_{1,1}$	$F_{1,2}$	$F_{1,3}$	$F_{1,4}$
1	108	1000	*4180*	—	=	—
2	124	892	3180	*8541*	*10848*	—
3	152	768	2288	5361		—
4	170	616	1520	2073	5487	*8995*
5	158	446	904	1553	2414	3508
6	139	288	458	649	861	1094
7	128	149	170	191	212	233
8	21	21	21	21	21	21
Σ	1000	4180	12721	—	—	—

Tem-se então os seguintes elementos:

$$h = 10$$

$$N = 1000$$

$$K = 115$$

$$F_{1,1} = 4180$$

$$F_{2,2} = 8541$$

$$F_{3,3} = 10848$$

$$F_{4,4} = 8995$$

4.6.1 — Cálculo dos Momentos Fatoriais

Aplicando-se a (17) tem-se os 4 momentos primeiros fatoriais:

$$\alpha_{(1)} = m = \frac{4\,180}{1\,000} = 4{,}18$$

$$\alpha_{(2)} = 2 \times \frac{8\,541}{1\,000} = 17{,}082$$

$$\alpha_{(3)} = 6 \times \frac{10\,848}{1\,000} = 65{,}088$$

$$\alpha_{(4)} = 24 \times \frac{8\,995}{1\,000} = 215{,}88$$

4.6.2 Cálculo dos Momentos Ordinários

Tendo em vista a (7) tem-se

$$\alpha_1 = \alpha_{(1)} = 4{,}18$$

$$\alpha_2 = \alpha_{(1)} + \alpha_{(2)} = 17{,}082 + 4{,}18 = 21{,}262$$

$$\alpha_3 = \alpha_{(1)} + \alpha_{(2)} + \alpha_{(3)} = 4{,}18 + 3x17{,}082 + 65{,}088 = 120{,}514$$

$$\alpha_4 = 4{,}18 + 7x17{,}082 + 6x65{,}088 + 215{,}88 = 730{,}162$$

Observe-se que esses mesmos resultados podem ser obtidos a partir dos $F_{i,k}$ através da (18).

4.6.3 Cálculo dos Momentos Centrais

Utilizando-se as relações (8) obtém-se:

$$\mu_2 = 21{,}262 - (4{,}18)^2 = 3{,}7896$$

$$\mu_3 = 120{,}514 - 3x21{,}262 \times 4{,}18 + 2x(4{,}18)^3 = -0{,}0422$$

$$\mu_4 = 730{,}162 - 3x120{,}514x4{,}18 + 6x21{,}262 \times \times (4{,}18)^2 - 3x(4{,}18)^4 = 28{,}3026$$

Observe-se que esses mesmos valores poderiam ser encontrados com a utilização dos momentos fatoriais. Para isto basta utilizar a (9), obtendo-se:

$$\mu_2 = 17{,}082 - 4{,}18\ (3{,}18) = 3{,}7896$$

$$\mu_3 = 65{,}088 - 3 \times 3{,}18 \times 17{,}082 + 4{,}18 \times 3{,}18 \times 7{,}36 = -0{,}0422$$

$$\mu_4 = 215{,}88 - 697{,}7436 + 1053{,}5221 - 543{,}3561 = 28{,}3026$$

que são valores encontrados anteriormente.

4.6.4 Cálculos dos Momentos Ordinários Corrigidos

Considerando-se a (22) tem-se

$$\alpha'_1 = \mu = 41{,}8 + 115 = 156{,}8$$

$$\alpha' = 1708{,}2 + 10032 + 13225 = 24965{,}2$$

$$a'_3 = 65088 + 640575 + 1806805 + 1520875 = 4043343$$

$$a'_4 = 2158800 + 33845760 + 160314570 + 289423200 + 174900615 = 660642955$$

Esses mesmos resultados podem ser encontrados empregando-se a (21).

4.6.5 Cálculo dos Momentos Centrais Corrigidos

Para o cálculo dos momentos centrais corrigidos basta multiplicar os valores de μ'_2, μ'_3, μ'_4, respectivamente por h^2, h^3, h^4, isto é, aplicando-se a (23):

$$\mu'_2 = 378{,}96$$

$$\mu'_3 = -42{,}2$$

$$\mu'_4 = 283026$$

5. PROCESSO ABREVIADO

Viu-se que para o emprego do método dos momentos fatoriais era preciso que a variável y assumisse os valores do conjunto $\{1, 2, \ldots, n\}$. Todavia, esta restrição não impede de utilizar um processo de simplificação dos cálculos através da utilização de uma variável reduzida x, que assuma os valores do conjunto $D = \{-p, -p+1, \ldots, -2, -1, 0, 1, 2, \ldots p-1, p\}$ onde $n = p + q + 1$. Então, se $X \varepsilon D$, x assumirá valores positivos e negativos, e é necessário reformular alguns conceitos já emitidos. Sejam então:

x^+ — Valores positivos de x

x^- — Valores negativos de x

f_i^+ — Freqüência simples da i-ésima classe dos valores de x^+

f_i^- — Freqüência simples da i-ésima classe dos valores de x^-

F_i^+ — Freqüência acumulada (maior que) da i-ésima classe dos valores de x^+

F_i^- — Freqüência acumulada (menor que) da i-ésima classe dos valores de x^-

$F_{i,k}^+$ — Freqüência acumulada de ordem k relativa a x^+

$F^{-}_{i,k}$ — Freqüência acumulada de ordem k relativa a x^{+}

$\alpha^{+}_{(k)}$ — Momento fatorial de ordem k relativo a x^{+}

$\alpha^{-}_{(k)}$ — Momentos fatorial de ordem k relativo a x^{-}

5.1 Momentos Ordinários

Então, por um raciocínio análogo ao desenvolvido nos itens (3.3) (4.1).

$$F^{-}_{k} = (-1)^{k} \frac{1}{k!} \sum_{j=k}^{p} j_{(k)} f_j \tag{24}$$

e que

$$\alpha^{-}_{(k)} = (-1)^{k} k! \frac{F^{-}_{k,k}}{N}$$

$$\alpha^{+}_{(k)} = k! \frac{F^{+}_{k,k}}{N} \tag{25}$$

observando-se que

$$\sum_{j=1}^{n} j f_j = \sum_{j=0}^{q} j f^{+}_j + \sum_{j=-1}^{-p} j f^{-}_j$$

$$\sum_{j=1}^{n} j^2 f_j = \sum_{j=0}^{q} j^2 f^{+}_j + \sum_{j=-1}^{-q} j^2 f_j =$$

$$\sum_{j=0}^{p} \{j_{(2)} + j_{(1)}\} f^{+}_j + \sum_{j=1}^{q} \{j_{(2)} + j_{(1)}\} f^{-}_j = \sum_{j=0}^{p} j_{(2)} f^{+}_j +$$

$$+ \sum_{j=1}^{q} j_{(2)} f^{-}_j + \sum_{j=0}^{p} j_{(1)} f^{+}_j + \sum_{j=1}^{q} j_{(1)} f^{-}_j$$

$$\sum_{j=1}^{n} j^3 f_j = \sum_{j=0}^{p} \{j_{(3)} + 3j_{(2)} + j_{(1)}\} f j^{+} + \sum_{j=1}^{-p} \{j_{(3)} + 3j_{(2)} + j_{(1)}\} f_j$$

$$\sum_{j=1}^{n} j^4 f_j = \sum_{j=0}^{p} \{j_{(4)} + 6\, j_{(3)} + 7\, j_{(2)} + j_{(1)}\} f^{+}_j +$$

$$\sum_{j=1}^{q} \{j_{(4)} + 6\, j_{(3)} + j_{(2)} + j_{(1)}\} f^{-}_j$$

Tendo em vista a (24) e a definição de momento ordinários, obtém-se:

$$\alpha_1 = \frac{F^+_{1,1} - F^-_{1,1}}{N}$$

$$\alpha_2 = \frac{2\ F^+_{2,2} + F^-_{2,2} + F^+_{1,1} + F^-_{1,1}}{N} \tag{26}$$

$$\alpha_3 = \frac{6\ F^+_{3,3} - F^-_{3,3} + 6\ F^+_{2,2} - F^-_{2,2} + F^+_{1,1} - F^-_{1,1}}{N}$$

$$\alpha_4 = \frac{24\ F^+_{4,4} + F^-_{4,4} + 36\ F^+_{3,3} + F^-_{3,3} + 14\ F^+_{2,2} + F^-_{2,2} + F^+_{1,1} + F^-_{1,1}}{N}$$

Substituindo-se as expressões de (25) em (26) obtém-se os momentos ordinários em função dos momentos fatoriais de x^+ e x^-, isto é:

$$\begin{aligned}
\alpha_1 &= \alpha^+_{(1)} + \alpha^-_{(1)} \\
\alpha_2 &= \left\{\alpha^+_{(2)} + \alpha^-_{(2)}\right\} + \left\{\alpha^+_{(1)} - \alpha^-_{(1)}\right\} \\
\alpha_3 &= \left\{\alpha^+_{(3)} + \alpha^-_{(3)}\right\} + \left\{\alpha^+_{(2)} - \alpha^-_{(2)}\right\} + \left\{\alpha^+_{(1)} + \alpha^-_{(1)}\right\} \\
\alpha_4 &= \left\{\alpha^+_{(4)} + \alpha^-_{(4)}\right\} + 6\left\{\alpha^+_{(3)} - \alpha^-_{(3)}\right\} + 7\left\{\alpha^+_{(2)} + \alpha^-_{(2)}\right\} + \left\{\alpha^+_{(1)} - \alpha^-_{(1)}\right\}
\end{aligned} \tag{27}$$

5.2 Tabela de Cálculos Auxiliares

Para o cálculo dos momentos pelo critério abreviado constrói-se a seguinte tabela:

x	f_i	F_i	$F_{i,1}$	$F_{i,2}$	$F_{i,3}$	$F_{i,4}$
—p	f^-_p	F^-_p	$F^-_{p,1}$	$F^-_{p,2}$	$F^-_{p,3}$	$F^-_{p,4}$
—p+1	f_{-p+1}	F_{-p+1}	$F_{-p+1,1}$	$F_{-p+1,2}$	$F_{-p+1,3}$	$F_{-p+1,4}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
—4	f^-_4	F^-_4	$F^-_{4,1}$	$F^-_{4,2}$	$F^-_{4,3}$	$F^-_{4,4}$
—3	f^-_3	F^-_3	$F^-_{3,1}$	$F^-_{3,2}$	$F^-_{3,3}$	—
—2	f^-_2	F^-_2	$F^-_{2,1}$	$F^-_{2,2}$	—	—
—1	f^-_1	F^-_1	$F^-_{1,1}$	—	—	—
0	f_0	F_0	—	—	—	—
1	f^+_1	F^+_1	$F^+_{1,1}$	—	—	—
2	f^+_2	F^+_2	$F^+_{2,1}$	$F^+_{2,2}$	—	—
3	f^+_3	F^+_3	$F^+_{3,1}$	$F^+_{3,2}$	$F^+_{3,3}$	—
4	f^+_4	F^+_4	$F^+_{4,1}$	$F^+_{4,2}$	$F^+_{4,3}$	$F^+_{4,4}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
q—1	f^+_{q-1}	F^+_{q-1}	$F^+_{q-1,2}$	$F^+_{q-1,2}$	$F^+_{q-1,3}$	$F^+_{q-1,4}$
q	f^+_q	F^+_q	$F^+_{q,1}$	$F^+_{q,2}$	$F^+_{q,3}$	$F^+_{q,4}$
Σ	N	—	—	—	—	—

5.3 Exemplo

Considerando-se a DF dada em (4.6) (vamos aplicar o processo abreviado para calcular os seus momentos. Construindo-se a tabela do item 5.1.3 obtém-se:

TABELA 2

$x_i = \frac{X_i - 165}{10}$	f_i	F_i	$F_{i,1}$	$F_{i,2}$	$F_{i,3}$	$F_{i,4}$
— 4	108	108	108	108	108	*108*
— 3	124	232	340	448	*556*	—
— 2	152	384	724	*1172*	—	—
— 1	170	554	*1278*	—	—	—
0	158	—	—	—	—	—
1	139	288	*458*	—	—	—
2	128	149	170	*191*	—	—
3	21	21	21	21	*21*	—
—	1000	—	—	—	—	—

Tem-se então:

$$h = 10$$

$$k = 165$$

$$F^{+}_{1,1} = 458 \qquad F^{-}_{1,1} = 1278$$

$$F^{+}_{2,2} = 191 \qquad F^{-}_{2,2} = 1172$$

$$F^{+}_{3,3} = 21 \qquad F^{-}_{3,3} = 556$$

$$F^{+}_{4,4} = 0 \qquad F^{-}_{4,4} = 108$$

5.3.1 *Cálculo dos Momentos Fatoriais*

Utilizando-se as expressões de (25) tem-se:

$$\alpha^{+}_{(1)} = \frac{458}{1\,000} = 0{,}458$$

$$\alpha^{-}_{(1)} = -\frac{1\,278}{1\,000} = -1{,}278$$

$$\alpha^{+}_{(2)} = 2 \times \frac{191}{1\,000} = 0{,}382$$

$$\alpha_{(2)}^{-} = 2 \times \frac{1\,172}{1\,000} = 2{,}344$$

$$\alpha_{(3)}^{+} = 6 \times \frac{21}{1\,000} = 0{,}126$$

$$\alpha_{(3)}^{-} = -\,6 \times \frac{556}{1\,000} = -\,3{,}336$$

$$\alpha_{(4)}^{+} = 24 \times \frac{0}{1\,000} = 0$$

$$\alpha_{(4)}^{-} = 24 \times \frac{108}{1\,000} = 2{,}592$$

5.3.2 *Momentos Ordinários*

Pela (27) os momentos ordinários são os seguintes:

$$\alpha_1 = 0{,}458 - 1{,}278 = -\,0{,}820$$

$$\alpha_2 = (0{,}382 + 2{,}344) + (0{,}458 + 1{,}278) = 4{,}462$$

$$\alpha_3 = (0{,}126 - 3{,}336) + (0{,}382 - 2{,}344) + (0{,}458 - 1{,}278) = 9{,}916$$

$$\alpha_4 = 2{,}592 + 6\,(0{,}126 + 2{,}344) + 7\,(0{,}382 + 2{,}344) + (0{,}458 - 1{,}278) = 44{,}182$$

Esses mesmos resultados podem ser obtidos através da (26), ou seja:

$$\alpha_1 = \frac{458 - 1278}{1000} = -\,0{,}820$$

$$\alpha_2 = \frac{2\,(191 + 1172) + (458 + 1278)}{1000} = 4{,}462$$

$$\alpha_3 = \frac{6\,(21 - 556) + 6\,(191 - 1172) + (458 - 1278)}{1000} = -\,9{,}916$$

$$\alpha_4 = \frac{24\,(0 + 108) + 36\,(21 + 556) + 14\,(191 + 1172) + (468 + 1278)}{1000} = 44{,}182$$

5.3.3 Momentos Centrais

Aplicando-se as relações (18) em

$$\mu_2 = 4{,}462 - (-0{,}820)^2 = 3{,}7896$$

$$\mu_3 = -9{,}916 - 3 \times (4{,}462)\ (-0{,}820) + 2\ (-0{,}820)^3 = \\ = -0{,}0422$$

$$\mu_4 = 44{,}182 - 4\ (-9{,}916)\ (-0{,}820) + 6\ (4{,}462) \\ (-0{,}820)^2 - 3\ (-0{,}820)^4 = 28{,}3026$$

que são os momentos encontrados anteriormente

5.3.4 Momentos Corrigidos

Neste caso tem-se

$$K = 165$$

$$h = 10$$

Aplicado a (21) em:

$$\mu = \alpha'_1 = -8{,}20 + 165 = 156{,}8$$

$$\alpha'_2 = 446{,}2 - 2706 + 27225 = 24965{,}2$$

$$\alpha'_3 = 9916 + 220869 - 669735 + 4492125 = 4033343$$

$$\alpha'_4 = 441820 - 6544560 + 72886770 - 147341700 + \\ + 741200625 = 660642955$$

que são os mesmos valores encontrados anteriormente. Para os momentos centrais corrigidos tem-se:

$$\mu'_2 = 100 \times 3{,}7896 = 378{,}96$$

$$\mu'_3 = 1000 \times (-0{,}0422) = -42{,}2$$

$$\mu'_4 = 10000 \times 28{,}3026 = 283026$$

Valores estes também já encontrados.

6. EFICIÊNCIA DO MÉTODO

A fim de comparar a eficiência do método dos momentos fatoriais, vamos calcular os momentos ordinários pelo processo usual para as variáveis y e x.

Para a variável y tem-se a seguinte tabela:

TABELA 3

Y_i	$Y_i = \frac{X_i - 115}{10}$	f_i	$y_i f_i$	$y^2_i f_i$	$y^3_i f_i$	$y^4_i f_i$
125	1	108	108	108	108	108
135	2	124	248	496	992	1984
145	3	152	456	1368	4104	12312
155	4	170	680	2720	10880	43520
165	5	158	790	3950	19750	98750
175	6	139	834	5004	30024	180144
185	7	128	896	6272	43904	307328
195	8	21	168	1344	10752	86016
Σ	—	1000	4180	21262	120514	730162

Tem-se, então, os momentos ordinários de y

$$\alpha_1 = \frac{4180}{1000} = 4{,}18$$

$$\alpha_2 = \frac{21262}{1000} = 21{,}262$$

$$\alpha_3 = \frac{120514}{1000} = 120{,}514$$

$$\alpha_4 = \frac{730162}{1000} = 730{,}162$$

Para a variável x a tabela de cálculos auxiliares é a seguinte:

TABELA 4

X_i	$x_i = \frac{X_i - 165}{10}$	f_i	$x_i f_i$	$x^2_i f_i$	$x^3_i f_i$	$x^4_i f_i$
125	— 4	108	— 432	1728	— 6912	27648
135	— 3	124	— 372	1116	— 3348	10044
145	— 2	152	— 304	608	— 1216	2432
155	— 1	170	— 170	170	— 170	170
165	0	158	—	—	—	—
175	1	139	139	139	139	139
185	2	128	256	512	1024	2048
195	3	21	63	189	567	1704
Σ	—	1000	— 820	4462	— 9916	44182

logo, os momentos ordinários de x são:

$$\alpha_1 = \frac{-820}{1000} = -0{,}820$$

$$\alpha_2 = \frac{4462}{1000} = 4{,}462$$

$$\alpha_3 = \frac{-9916}{1000} = -9{,}916$$

$$\alpha_4 = \frac{44182}{1000} = 44{,}182$$

como se vê, da comparação das tabelas (1) com (3) ou das tabelas (2) e (4) existe uma vantagem para o método dos momentos, porque as tabelas de cálculos auxiliares são menos trabalhosas e cujos números apresentam menor ordem de grandeza.

ANÁLISE ESTATÍSTICA DO PODER DISCRIMINATIVO DE QUESTÕES DE PROVAS

Hervey Guimarães Cova *

SUMÁRIO

* Professor da Escola Nacional de Ciências Estatísticas — ENCE/IBGE.

1. INTRODUÇÃO

O processo seletivo que vem sendo aplicado aos candidatos que se inscrevem nos exames vestibulares para ingresso nas universidades brasileiras tem oferecido campo a muitas discussões e controvérsias.

Mesmo sem entrar numa análise mais profunda dos diversos aspectos que o problema envolve, cabe sejam formuladas as seguintes perguntas:

1.ª — O processo seletivo em vigor está, de fato, selecionando os mais aptos aos diversos cursos?

2.ª — Até que ponto cada questão de prova já realizada contribuiu, positivamente, para emprestar validade ao processo seletivo no qual se inseriu?

Em matéria de exames vestibulares, o critério de seleção geralmente adotado funda-se na hipótese — embora não declarada — segundo a qual cada questão acertada por um candidato aumenta a probabilidade de que o mesmo possua a aptidão necessária para realizar, com êxito, o curso por ele visado.

Entretanto, essa hipótese não vem sendo submetida sistematicamente a algum tipo de teste — realizado *a posteriori* — que a confirme ou rejeite. Outrossim, se atentarmos para o fundamento estatístico em que se apóia o processo científico de elaboração de certos instrumentos de medidas psicopedagógicas, veremos que a referida hipótese não pode ser aceita pacificamente.

De fato, seja um grupo numeroso, A, de indivíduos possuindo algumas características em comum; e seja outro grupo numeroso, B, de indivíduos não apresentando essas características. Suponhamos ainda que um pesquisador — no campo da Psicologia ou da Pedagogia — tivesse constatado, através de um grande número de observações, que os indivíduos do grupo A, diante de certo estímulo E, costumam apresentar a resposta R_1; ao passo que, diante do mesmo estímulo E, os indivíduos do grupo B costumam apresentar a resposta R_2. Então, ao se defrontar, posteriormente, com algum outro indivíduo, que, diante do estímulo E, oferecesse a resposta R_1, tal pesquisador — com base em seus conhecimentos científicos — aceitaria como bastante provável a hipótese de que este último indivíduo possua as características do grupo A.

Transportemos a situação anterior para a prática dos exames vestibulares, colocando uma nova situação na qual o grupo A seja constituído pelos candidatos que possuem, num grau mais elevado, a aptidão para realizar com êxito um determinado curso; e o grupo B seja formado pelos candidatos que não possuem (ou que a possuem num grau bem mais baixo) a referida aptidão. Substituamos ainda o estímulo E por

alguma questão de prova Q. Por último, admitamos que os candidatos do grupo A, de um modo geral, diante da questão Q, apresentem a resposta R_1 e que os candidatos do grupo B, diante da mesma questão, apresentem a resposta R_2. Naturalmente, devemos acrescentar que uma das respostas — R_1 ou R_2 — será a *resposta certa* à questão Q; a outra, portanto, devendo representar, genericamente, qualquer resposta errada, à dita questão.

Embora esta nova situação seja diferente, em seus aspectos exteriores, da situação inicialmente focalizada, ambas apresentam o mesmo conteúdo intrínseco; de sorte que um juiz imparcial, quer analisando as respostas ao estímulo E quer analisando as respostas à questão Q, deve assumir, diante dos fatos, atitudes coerentes.

Ora, uma possível falha do processo seletivo vigente (hipótese que estamos propondo seja testada) reside em se admitir que R_1 (resposta típica dos candidatos do grupo A) seja sempre a *resposta certa* à questão Q; e que R_2 (resposta típica dos candidatos do grupo B) seja sempre a genérica resposta errada à mesma questão Q; quando o contrário, em verdade, poderia estar ocorrendo.

Entendemos, pois, que um examinador criterioso, desejando tomar uma *atitude científica* diante dos fatos deve ter o cuidado — a fim de evitar a repetição, no futuro, de erros presentes — de investigar se a resposta típica dos candidatos do grupo A foi, de fato, a resposta certa à questão Q; e se a resposta típica dos candidatos do grupo B foi, de fato, a genérica resposta errada a tal questão.

Se os examinadores que integram as bancas dos exames vestibulares tivessem a preocupação de analisar, *a posteriori,* o valor discriminativo das questões de provas, através dos resultados obtidos pelos candidatos que se submeteram a tais provas, certamente reconheceriam a existência de questões que não teriam contribuído, de qualquer forma, para a seleção dos mais aptos; podendo até identificar, em casos mais raros (porém não impossíveis), a existência de questões que teriam beneficiado, no processo seletivo, os candidatos menos aptos, em detrimento dos mais aptos.

Num primeiro instante pode parecer estranha a possibilidade de alguma questão de prova favorecer a seleção dos menos aptos. Entretanto, diversos fatores, nem sempre conhecidos ou controláveis, poderão contribuir para isso.

Imaginemos, num processo seletivo, a inclusão de uma *prova de conhecimentos gerais* da qual constassem perguntas tais como: "em que partida de futebol Pelé marcou seu milésimo gol?", ou, ainda, "qual o nome do autor da música classificada em primeiro lugar, no segundo festival de música popular brasileira?". Ora, podemos conceber, de um lado, um tipo de candidato, X, sempre voltado para os trabalhos escolares e para o estudo de assuntos sérios, que, por isso mesmo, dedicasse

uma parcela mínima de seu tempo a acompanhar e discutir fatos relacionados com futebol e música popular; e, de outro lado, um tipo de candidato, Y, pouco amigo dos estudos e dos assuntos sérios, mas que, em compensação, dedicasse a maior parte de seu tempo a acompanhar e discutir fatos daquela natureza. Evidentemente, perguntas como as que foram acima formuladas tenderiam a favorecer, com predominância, aos candidatos do tipo Y, entre os quais poderiam estar, em maioria, os menos aptos.

Portanto, um trabalho sistemático de análise das questões de provas haveria de contribuir, eficazmente, para o aperfeiçoamento do processo seletivo, através da eliminação futura das questões que não possuem poder discriminativo, ou que o possuem no sentido negativo.

A avaliação do poder discriminativo das questões de provas, mediante a análise posterior de seus resultados, é um problema que tem sido abordado mais de uma vez em obras versando sobre elaboração ou utilização de instrumentos de medida, em educação. Entretanto, parece que o assunto, na prática, não vem sendo tratado com a importância que realmente merece.

Essa falta de interesse poderia ser, talvez, uma decorrência do fato de não ter sido proposto, até o presente, um critério mais rigoroso — em seu aspecto matemático probabilístico — para a análise e classificação das questões de provas, segundo o ponto de vista que estamos aqui discutindo.

A presente monografia tem, assim, dois objetivos:

1.º — estabelecer, mediante um desenvolvimento teórico adequado, um critério prático, objetivo e uniforme para a avaliação das questões de provas, de acordo com seu poder discriminativo;

2.º — aplicar esse critério a uma situação real, a fim de exemplificar o seu funcionamento na prática.

Aos dois objetivos acima declarados estão dedicadas, respectivamente, a segunda e a terceira partes do nosso trabalho.

2. PROPOSTA DE UM CRITÉRIO DE CLASSIFICAÇÃO DE QUESTÕES DE PROVAS (fundamentação teórica)

2.1 Colocação do problema

Imaginemos que um conjunto de N candidatos a um determinado curso deva ser submetido a uma prova, cuja finalidade seja a de selecioná-los em dois grupos, A e B, de tal sorte que os candidatos incluídos no grupo A possam ser considerados nitidamente superiores aos incluí-

dos no grupo B, no que se refere à aptidão para realizar com êxito o referido curso.

Admitamos ainda que a suposta prova conste de várias questões e que o critério de seleção esteja baseado no total de pontos obtido pelo candidato nas questões que acertar, de forma que, dentre dois candidatos, seja considerado o mais apto aquele que obtiver maior número de pontos na citada prova.

Suponhamos, finalmente, que, após a realização da prova e a separação dos candidatos nos dois grupos referidos ao início, tenha-se chegado ao seguinte quadro resumo, abrangendo os resultados parciais, relativamente a uma dada questão, Q:

QUADRO 1

RESULTADOS DA QUESTÃO Q

GRUPO	NÚMERO DE CANDIDATOS DO GRUPO QUE ACERTARAM A QUESTÃO Q	NÚMERO DE CANDIDATOS INCLUÍDOS NO GRUPO
A	N_{AC}	N_A
B	N_{BC}	N_B
Total de candidatos	N_C	N

Preliminarmente, consideremos as razões

$$R_1 = \frac{N_{AC}}{N_A} \qquad \text{e} \qquad R_2 = \frac{N_{BC}}{N_B}$$

que indicam, respectivamente, a proporção de candidatos dentro do grupo A e dentro do grupo B que acertaram a questão Q.

Se ficar constatado que a questão Q foi *muito mais* acertada (em termos relativos) pelo grupo A do que pelo grupo B — ou seja, se R_1 for muito maior que R_2 — o bom senso nos dirá que tal questão, *quando acertada,* deve traduzir alguma qualidade que é mais característica de candidatos do grupo A.

Ao contrário, se a questão Q foi muito mais acertada pelo grupo B do que pelo grupo A — isto é, se R_1 for muito menor que R_2 — tal questão, *quando acertada,* deve traduzir alguma qualidade que é mais característica de candidatos do grupo B.

Finalmente, se a questão Q foi *igualmente acertada* (em termos relativos) pelos dois grupos — vale dizer, se os valores de R_1 e R_2 forem

aproximadamente iguais — essa questão nada pode revelar, objetivamente, sobre alguma qualidade característica de um ou de outro grupo.

Essas considerações preliminares não constituem nenhuma novidade. Representam, como já declaramos na introdução, o fundamento estatístico das técnicas de elaboração de certos tipos de testes (instrumentos de medidas psicopedagógicas), para as mais diversas finalidades.

Prosseguindo, pois, em nossa linha de raciocínio, admitamos que a classificação dos candidatos nos dois grupos, A e B, feita com base nos resultados da prova, esteja correta, isto é, admitamos que os candidatos do grupo A sejam, de fato, os mais aptos a seguirem o curso. Tal hipótese equivale a aceitar que a prova supostamente realizada foi um instrumento válido para medir aquela aptidão. Por conseguinte, se um candidato qualquer fosse escolhido ao acaso, do grupo inicial e, posteriormente, tivéssemos a informação de que o referido candidato acertou a questão Q, esse conhecimento parcial de seu desempenho deveria *aumentar* a probabilidade de que ele tenha sido classificado no grupo A (grupo superior). Se isto não acontecer, ou melhor, se aquela informação parcial de que o candidato acertou a questão Q acarretar a *diminuição* da probabilidade de que ele tenha sido classificado no grupo A, seremos levados a concluir, mediante raciocínio inverso, que a questão Q não se presta ao diagnóstico da aptidão dos candidatos relativamente ao curso por eles visado.

Nessas condições, para decidir da validade da prova como um todo, devemos analisar os resultados de cada questão em particular, a fim de *medir* até que ponto cada uma delas estaria contribuindo para a correta classificação dos candidatos.

2.2 Classificação genérica das questões

Com tal objetivo e voltando a admitir que um candidato tenha sido escolhido *ao acaso* do grupo inicial dos N candidatos, consideremos os seguintes eventos:

A = {o candidato escolhido foi classificado no grupo A}

C = {o candidato escolhido acertou a questão Q}

Desde que a escolha do candidato seja feita ao acaso, a probabilidade de que ele tenha sido classificado no grupo A — de acordo com o quadro resumo dos resultados da questão Q — será dada por

$$P(A) = \frac{N_A}{N} \qquad (1)$$

Entretanto, se tivermos a informação de que tal candidato acertou a questão Q — certeza da ocorrência do evento C — a probabilidade de que ele tenha sido classificado no grupo A deverá ser agora recalculada mediante:

$$P(A \mid C) = \frac{N_{AC}}{N_C} \tag{2}$$

Ora, combinando (1) e (2), resulta:

$$\frac{P(A \mid C)}{P(A)} = \frac{\frac{N_{AC}}{N_C}}{\frac{N_A}{N}} = \frac{N_{AC}}{N_C} \cdot \frac{N}{N_A} = \frac{N_{AC}}{N_A} \cdot \frac{N}{N_C} = \frac{N_{AC}}{N_A} \Big/ \frac{N_C}{N};$$

e, introduzindo as razões:

$$R_1 = \frac{N_{AC}}{N_A} \qquad \text{e} \qquad R_3 = \frac{N_C}{N},$$

obtemos, finalmente:

$$\boxed{P(A \mid C) = P(A) \cdot \frac{R_1}{R_3}}, \tag{3}$$

onde R_1, como já vimos antes, traduz a proporção de candidatos — dentro do grupo A — que acertaram a questão Q; e R_3 exprime a proporção geral de candidatos — dentro dos dois grupos — que acertaram a mesma questão.

Vamos agora analisar a relação (3) mediante as três hipóteses seguintes:

1.ª Hipótese: $R_1 = R_3$.

Nesse caso resulta: $P(A \mid C) = P(A)$, isto é: a informação de que o candidato acertou a questão Q não afeta a probabilidade que ele tenha sido incluído no grupo A.

Diremos, por isso mesmo, que a questão Q é *não discriminativa.*

2.ª Hipótese: $R_1 > R_3$.

Agora obtemos: $P(A) \mid C) > P(A)$, vale dizer: a informação de que o candidato acertou a questão Q aumenta a probabilidade de que ele tenha sido classificado no grupo A.

Diremos, então, que a questão Q é *discriminativa positiva.*

3.ª Hipótese: $R_1 < R_3$.

Nesta última hipótese achamos: $P(A \,/\, C) < P(A)$, ou ainda: a informação de que o candidato acertou a questão Q diminui a probabilidade de que ele tenha sido classificado no grupo A.

Diremos agora que a questão Q é *discriminativa negativa.*

Confirmamos, assim, que uma questão de prova pode revelar-se de três maneiras diferentes, em face da probabilidade de um candidato — que a tenha acertado — haver sido incluído no grupo superior A: não afetando essa probabilidade, ou aumentando-a, ou diminuindo-a; conforme já havíamos previsto.

Acontece, todavia, que dificilmente ocorrerá, na prática, a primeira hipótese: $R_1 = R_3$; por isso mesmo, se não adotarmos um critério mais racional para a classificação das questões, todas elas serão classificadas como discriminativas, positivas ou negativas; sendo possível a ocorrência de casos em que duas questões, apresentando valores muito próximos para o quociente R_1/R_3, tenham de ser classificadas em campos opostos.

É necessário, pois, introduzir um critério que permita classificar como "não discriminativas" as questões que fornecerem o quociente R_1/R_3 aproximadamente igual à unidade; e que possibilite ainda estabelecer uma gradação de intensidade para as questões classificáveis como "discriminativas".

2.3 Novo critério de classificação genérica das questões

Antes de abordar o problema principal que nos propomos — fixação de um critério prático, objetivo e uniforme de classificação de questões de provas — vamos estabelecer um novo critério de classificação genérica de tais questões:

Para isso, introduziremos a nova razão

$$k = \frac{N_B}{N_A},$$

da qual deduzimos $N_B = k\, N_A$ e, daí:

$$N = N_A + N_B = N_A\ (1 + k)$$

Portanto:

$$\frac{R_1}{R_3} = \frac{N_{AC}}{N_A} \cdot \frac{N}{N_C} = \frac{N_{AC}}{N_A} = \frac{N_A\ (1 + k) = N_{AC}\ (1 + k)}{N_C} \qquad (4)$$

Outrossim, entrando com a relação

$$r = \frac{R_1}{R_2} = \frac{N_{AC}}{N_A} \cdot \frac{N_B}{N_{BC}} = \frac{N_{AC}\ k\ N_A}{N_A\ N_{BC}} = \frac{k\ N_{AC}}{N_{BC}},$$

obtemos:

$$N_{BC} = \frac{k\ N_{AC}}{r}$$

Nessas condições, resulta:

$$N_C = N_{AC} + N_{BC} = N_{AC}\ (1 + \left(\frac{k}{r}\right) \tag{5}$$

Combinando a (4) com a (5), achamos, finalmente:

$$\boxed{\frac{R_1}{R_3} = \frac{1 + k}{1 + \frac{k}{r}}} \tag{6}$$

Devemos notar que k, representando a razão entre os números N_B e N_A, de candidatos incluídos, respectivamente, no grupo B e no grupo A, não depende, geralmente, dos resultados particulares de cada questão (na prática, quando estamos diante de um processo puramente seletivo, N_B e N_A dependem apenas do total de candidatos inscritos e da disponibilidade de vagas para o curso); ao passo que r assume valores maiores ou menores de acordo com os resultados de cada questão em particular. Vamos, pois, que a relação R_1/R_3 depende, em última análise, do comportamento de $r = R_1/R_2$.

Se tivermos $R_1 = R_2$ — isto é, $r = 1$ — a relação (6) fornecerá $R_1 = R_2$. Logo, $r = 1$ revela tratar-se de questão "não discriminativa".

Se tivermos $R_1 > R_2$ — ou seja, $r > 1$ — a mesma relação (6) dará $R_1 > R_3$. Portanto, $r > 1$ indica tratar-se de questão "discriminativa positiva".

Se tivermos, finalmente, $R_1 < R_2$ — vale dizer, $r < 1$ — ainda a relação (6) acarretará $R_1 < R_3$. Assim, $r < 1$ informa tratar-se de questão "discriminativa negativa."

Acabamos de ver, pois, que a classificação das questões tanto poderá ser feita pelo comportamento de R_1/R_3, como poderá ser feita mediante o comportamento de $r = R_1/R_2$.

2.4 A necessidade de uniformizar o critério de classificação

Nosso próximo passo, no sentido de alcançar o objetivo visado inicialmente, será definir certos intervalos de variação de r, de tal sorte que, sempre que duas ou mais questões fornecerem valores de r dentro do mesmo intervalo, tais questões recebam a mesma classificação.

Aqui defrontamo-nos com uma séria dificuldade: tendo presentes as relações

$$P(A \mid C) = P(A)\,\frac{R_1}{R_3} \qquad \text{e} \qquad \frac{R_1}{R_3} = \frac{1+k}{1+\dfrac{k}{r}},$$

é fácil constatar que um mesmo valor de r poderá fornecer diferentes valores para o quociente R_1/R_3 — dependendo do correspondente valor de k —; e isto pode conduzir a uma falsa idéia do poder discriminativo (positivo ou negativo) das questões de prova sob análise. Apenas a título de exemplo, apresentamos o quadro abaixo, onde estão consignados, para um mesmo valor de r, diversos valores de R_1/R_3, em correspondência com os respectivos valores de k:

Para r = 2,00

k	R_1/R_3
0,5	1,20
1,0	1,33
2,0	1,50
5,0	1,71

Conforme vemos nesse quadro, se tivermos $k = 0{,}5$, o valor de $r = 2{,}00$ fornecerá $R_1/R_3 = 1{,}20$; significando isto que o fato de se ter a informação de que um candidato — escolhido ao acaso do grupo inicial — acertou a questão Q aumentaria em 20% a probabilidade de tal candidato ter sido incluído no grupo superior, A. Entretanto, se tivermos $k = 5{,}0$, o mesmo valor $r = 2{,}0$ fornecerá $R_1/R_3 = 1{,}71$; vale dizer: a informação de que o referido candidato acertou a questão Q aumentaria em 71% a probabilidade de ele ter sido incluído no grupo A.

Na primeira situação focalizada poderíamos ser tentados a afirmar que a questão Q possui um *fraco* poder discriminativo positivo; ao passo que, na segunda situação, talvez fôssemos induzidos a declarar que a questão Q possui um *forte* poder discriminativo positivo. Ora, como em qualquer das duas situações focalizadas, partimos de um mesmo valor $r = 2{,}0$; e como, por outro lado, a classificação das questões, conforme

já vimos antes, poderia ser feita também através dos valores de r, as supostas classificações, antes referidas, não teriam obedecido a um critério uniforme.

Para contornar essa dificuldade precisamos aprofundar nosso conhecimento sobre a relação (6).

2.5 Escolha de uma curva padrão

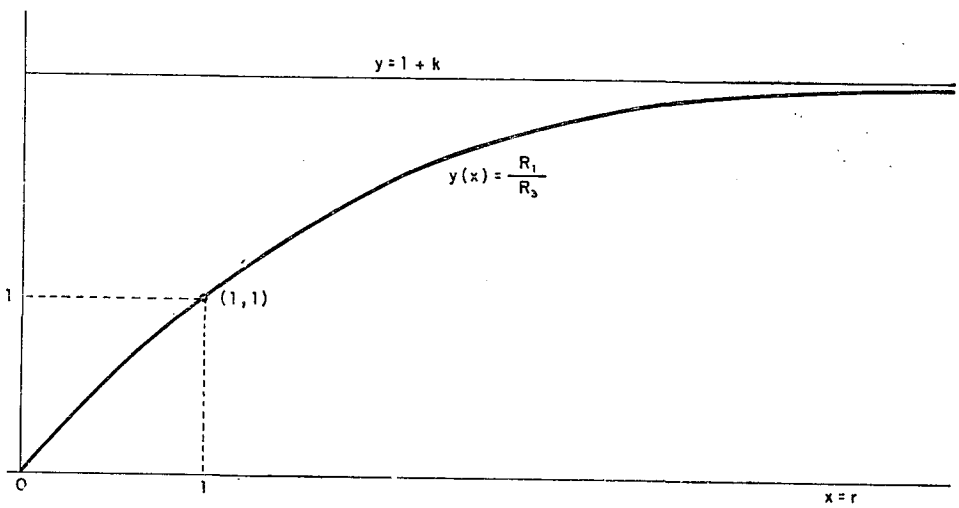
É fácil constatar que

$$\lim_{r \to 0} \frac{R_1}{R_3} = 0 \qquad \text{e} \qquad \lim_{r \to \infty} \frac{R_1}{R_3} = 1 + k$$

Outrossim, se fizermos, por um momento, $r = x$ e $y(x) = R_1/R_3$, a função $y(x)$ representará uma família de curvas, dependentes do parâmetro k. Entretanto, para qualquer valor desse parâmetro, a curva correspondente passará pelo ponto de coordenadas (1; 1) e será assintótica à reta de equação $y = 1 + k$. Além disso, sendo

$$y'(x) = k(1 + k)(k + x)^{-2} \text{ e } y''(x) = -2k(1 + k)(k + x)^{-3}$$

resulta que, para todo valor de $x > 0$, $y'(x) > 0$ e $y''(x) < 0$; de onde concluímos que $y(x)$ é uma função monótona crescente e que a curva por ela representada tem sua concavidade voltada para o eixo dos x, conforme mostramos a seguir:



DILUS/S. 01 - R.C.N.

Retomando o exame da relação (3), parece-nos razoável denominar o fator R_1/R_3 de *fator de incremento;* por ser tal fator o responsável pelas alterações ocorridas na probabilidade do evento A, quando se tem a informação da ocorrência do evento C.

Isto posto, estudando o comportamento das curvas da família acima referida, percebemos, claramente, que as variações ocorridas nos valores de $r = R_1/R_2$ — a partir do ponto de equilíbrio $r = 1$ — produzem *variações cada vez mais rápidas* no fator de incremento *à medida em que o valor do parâmetro k aumenta.*

Assim — vide tabela apresentada no subtítulo 2.4 — quando r cresce de 1,00 até 2,00: o fator de incremento cresce de 1,00 até 1,20, se $k = 0{,}5$; cresce de 1,00 até 1,50, se $k = 2$; e cresce de 1,00 até 1,71, se $k = 5$.

Analogamente, podemos ver que, quando r decresce de 1,00 até 0,50: o fator de incremento decresce de 1,00 até 0,75, se $k = 0{,}5$; decresce de 1,00 até 0,60, se $k = 2$; e decresce de 1,00 até 0,54, se $k = 5$.

Ora, quanto *mais rápido* a variação do fator de incremento tanto *mais fácil* será produzi-la; logo, como tais variações se transmitem no mesmo sentido à probabilidade do evento A, torna-se cada vez mais fácil aumentar ou diminuir essa probabilidade à medida em que o valor do parâmetro k aumenta.

Interpretando da forma acima as referidas variações, somos levados a concluir que um aumento de 20% na probabilidade do evento A, quando $k = 0{,}5$, deve ser equivalente a um aumento de 50%, na referida probabilidade, quando $k = 2$; e deve ser equivalente ainda a um aumento de 71% na mesma probabilidade, quando $k = 5$. Conclusões semelhantes seriam válidas se considerássemos os incrementos negativos.

De um modo geral, portanto, tomaremos como *equivalentes* todas as variações do fator de incremento que correspondam ao mesmo valor de r, *independentemente dos valores do parâmetro k,* isto é, independentemente das curvas da referida família que traduzam aquelas variações.

Tudo isso conduz à idéia de se tomar uma das curvas da família como *curva padrão,* através da qual seriam fixados alguns pontos de referência no eixo dos x; e, portanto, certos intervalos de variação de r que permitam classificar, uniformemente, as questões de prova.

Embora, de certo modo, essa escolha possa ser arbitrária, preferimos tomar como padrão a curva de equação

$$\frac{R_1}{R_3} = \frac{2}{1 + \frac{1}{r}},$$

correspondente ao valor $k = 1$, do parâmetro k, e que fornece para R_1/R_3 valores pertencentes ao intervalo aberto (0,2).

Essa preferência foi motivada pelos seguintes fatos:

1.º) Quando $k = 1$, o ponto de coordenadas (1, 1) — característico das questões não discriminativas — tem para a ordenada o ponto médio do intervalo (0, 2) que é, como assinalamos, o intervalo de variação de R_1/R_3; havendo, assim, a possibilidade de decompor o dito intervalo em uma escala uniforme e simétrica;

2.º) Os valores de r, correspondentes à curva de parâmetro $k = 1$, são aqueles que apresentam maior estabilidade, quando pensamos em termos de *pequenas variações ocasionais* que possam afetar as razões R_1 e R_2.

De fato, apenas para fixar idéias, imaginemos uma situação de prova na qual — para $k = 10$ — tivéssemos $N_A = 20$ e $N_B = 200$; e suponhamos que os resultados da questão Q, de tal prova, consignassem para os grupos A e B, respectivamente, as freqüências de acertos $N_{AC} = 5$ e $N_{BC} = 30$. Para esses valores fictícios, resultariam as razões

$$R_1 = \frac{N_{AC}}{N_A} = \frac{25}{100} \quad \text{e} \quad R_2 = \frac{N_{BC}}{N_B} = \frac{15}{100},$$

as quais forneceriam $r = 25/15 = 1{,}667$.

Ora, admitamos que, por mero acaso, mais um candidato, em cada grupo, acertasse a questão Q, dando origem às novas freqüências de acertos $N_{AC} = 6$ e $N_{BC} = 31$. Teríamos, então, as novas razões

$$R_1 = \frac{30}{100} \quad \text{e} \quad R_2 = \frac{15{,}5}{100},$$

e, a partir destas, o novo valor $r = 30/15{,}5 = 1{,}940$, muito mais elevado que o primeiro.

Vemos, por este único exemplo, que, sendo um dos valores N_A ou N_B muito pequeno e o outro muito grande (em termos relativos), uma das razões, R_1 ou R_2, será fortemente influenciada, mesmo por ligeiras variações (flutuações de amostra) ocorridas nas freqüências N_{AC} e N_{BC}; ao passo que a outra razão quase não será afetada por tais variações; de sorte que, em última análise, o quociente $r = R_1/R_2$ será sensivelmente modificado.

Nessas condições, confirmando o que dissemos ao início, a situação mais favorável à estabilidade dos valores de r é aquela em que $N_A = N_B$, isto é, na qual temos $k = 1$.

2.6 Passagem da curva observada para a curva padrão

Voltemos a considerar uma situação de prova (supostamente real) que nos fornecesse o quadro resumo *Q-1*, introduzido no subtítulo 2.1, onde estariam consignados os resultados parciais relativos à questão *Q*.

Partindo daqueles resultados parciais — que, por hipótese, teriam sido observados — chegamos à curva de equação

$$\frac{R_1}{R_2} = \frac{1 + k}{1 + \frac{k}{r}}, \qquad (1)$$

à qual, por isso mesmo, chamaremos de *curva observada.*

Ora, a cada situação real, em que $N_A \neq N_B$, podemos fazer corresponder uma situação ideal em que $N_A = N_B$.

De fato, supondo, por exemplo, $N_A < N_B$ (caso mais freqüente), se multiplicarmos os valores de N_{AC} e de N_A pelo fator $p = 1/k$, o quadro *Q-1*, resumindo a situação real, toma a forma:

QUADRO 2

RESULTADOS (IDEAIS) DA QUESTÃO Q

GRUPO	NÚMERO DE CANDIDATOS, DO GRUPO, QUE ACERTARAM A QUESTÃO Q	NÚMERO DE CANDIDATOS INCLUÍDOS NO GRUPO
A'	$\rho\ N_{ac}$	$\rho\ N_A$
B	N_{bc}	N_B
Total de candidatos	N'_c	N'

Esses resultados (ideais) fornecem os novos valores:

$$k' = \frac{\rho N_A}{N_B} = \frac{1}{k} \cdot k = 1 \quad ,$$

$$R'_1 = \frac{\rho N_{AC}}{\rho N_A} = \frac{N_{AC}}{N_A} = R_1 \quad , \quad R'_2 = \frac{N_{BC}}{N_B} = R_2 \quad \text{e} \quad R'_3 = \frac{N'_C}{N'}$$

dos quais deduzimos:

$$r' = \frac{R'_1}{R'_2} = \frac{R_1}{R_2} = r$$

Nessas condições, os resultados (ideais) do quadro *Q-2* nos levam à curva padrão

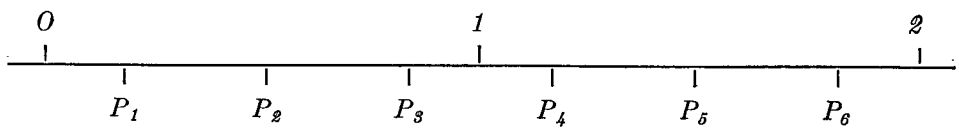
$$\frac{R_1}{R_3'} = \frac{2}{1 + \frac{1}{r}} \tag{2}$$

conforme havíamos anunciado.

Ora, examinando as relações (1) e (2), concluímos: o valor de r, que na curva observada fornece o fator de incremento R_1/R_3, *é o mesmo valor de r que, na curva padrão,* produz o fator de incremento R_1/R_3. Portanto, tendo em vista estarmos aceitando como equivalentes os valores do fator de incremento correspondentes ao mesmo valor de r, ficou confirmada, assim, a possibilidade de serem classificados, de imediato, os valores de r de qualquer curva observada (logo, a possibilidade de serem classificadas as próprias questões de prova), desde que já esteja disponível uma escala adequada para a classificação dos valores de r relativos à curva padrão.

2.7 Definição de uma escala de classificação

Para construir a escala desejada, convém lembrar que, na curva padrão, quando r percorre o semi-eixo $(0, \infty)$, o fator de incremento, R_1/R_3, varia de 0 (zero) a 2 (dois). Então — dando um primeiro passo no sentido de obter a referida escala — propomos seja dividido o intervalo (0, 2) em sete subintervalos, mediante os pontos: $P_1 = 0{,}2857$; $P_2 = 0{,}5714$; $P_3 = 0{,}8571$; $P_4 = 1{,}1428$; $P_5 = 1{,}4285$ e $P_6 = 1{,}7142$. Obtemos, assim, a seguinte escala — para a curva padrão — relativa aos valores de R_1/R_3:



De acordo com essa primeira escala e levando em conta os respectivos valores do fator de incremento — referidos à curva padrão — as questões de prova poderão ser classificadas do seguinte modo:

Fortemente discriminativa negativa — se $R_1/R_3 \in (0, P_1)$;
Moderadamente discriminativa negativa — se $R_1/R_3 \in (P_1, P_2)$;
Fracamente discriminativa negativa — se $R_1/R_3 \in (P_2, P_3)$;
Não discriminativa — se $R_1/R_3 \in (P_3, P_4)$;
Fracamente discriminativa positiva — se $R_1/R_3 \in (P_4, P_5)$;
Moderadamente discriminativa positiva — se $R_1/R_3 \in (P_5, P_6)$;
Fortemente discriminativa positiva — se $R_1/R_3 \in (P_6, 2)$.

Entretanto, a classificação das questões, tomando por base os valores do fator de incremento, não convém às aplicações práticas, pelo fato de exigir o trabalho preliminar de transformação da curva observada (quadro Q-1) na curva padrão (quadro Q-2). Para contornar esse inconveniente precisamos transformar a escala referida aos valores de R_1/R_3 em outra escala referida aos valores de r; bastando, para isso, decompor o semi-eixo $(0, \infty)$ em sete subintervalos, mediante a localização dos pontos desse semi-eixo que correspondem aos pontos P_1, P_2, P_3, P_4, P_5 e P_6, anteriormente considerados. Indicando os novos pontos por P'_1, P'_2, P'_3, P'_4, P'_5 e P'_6, a equação (1), do subtítulo anterior, fornece:

$$P_i = \frac{2}{1 + \frac{1}{P'_i}}$$

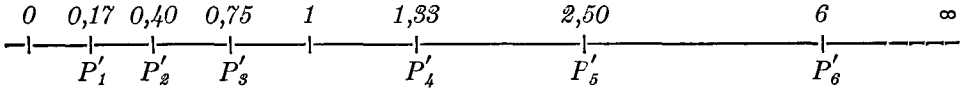
de onde deduzimos

$$P'_i = \frac{1}{\frac{2}{P_i} - 1}$$

Substituindo os P_i pelos valores já conhecidos, achamos (arredondando os resultados até os centésimos):

$$P'_1 = 0{,}17 \qquad P'_2 = 0{,}40 \qquad P'_3 = 0{,}75$$
$$P'_4 = 1{,}33 \qquad P'_5 = 2{,}50 \qquad P'_6 = 6{,}00$$

Chegamos, assim, à escala definitiva, referida aos valores de r:



onde os intervalos $(0, P'_1), (P'_1, P'_2), (P'_2, P'_3), (P'_3, P'_4), (P'_4, P'_5)$ (P'_5, P'_6) e (P, P'_∞), guardando a mesma ordem de correspondência, permitem a classificação das questões, conforme proposto anteriormente, mediante o cálculo direto dos respectivos valores de r.

2.8 Forma prática de utilização da escala

A classificação das questões, utilizando a nova escala, poderá ser feita com a maior facilidade, mesmo por professores que não estejam familiarizados com cálculos matemáticos ou estatísticos:

Uma vez conhecidos os valores N_A e N_B e obtidas as freqüências de acertos N_{AC} e N_{BC}, referentes à questão Q, calculamos as razões

$$R_1 = \frac{N_{AC}}{N_A} \quad \text{e} \quad R_2 = \frac{N_{BC}}{N_B}$$

que conduzem ao valor de r:

$$r = \frac{R_1}{R_2}$$

Em seguida, consultando o quadro abaixo, identificamos o intervalo em que está localizado esse valor de r, obtendo, logo à direita, a classificação da questão.

QUADRO 3

ESCALA PRÁTICA DE CLASSIFICAÇÃO DE QUESTÕES

INTERVALO DE LOCALIZAÇÃO DE r	CLASSIFICAÇÃO DA QUESTÃO	ABREVIATURA
0,00 — 0,17	Fortemente discriminativa negativa	DNF
0,17 ⊣ 0,40	Moderadamente discriminativa negativa	DNM
0,40 ⊣ 0,75	Fracamente discriminativa negativa	DNf
0,75 ⊣ 1,33	Não discriminativa	ND
1,33 ⊣ 2,50	Fracamente discriminativa positiva	DPf
2,50 ⊣ 6,00	Moderadamente discriminativa positiva	DPM
6,00 — ∞	Fortemente discriminativa positiva	DPF

3. APLICAÇÃO A UMA SITUAÇÃO REAL

3.1 Escolha da situação real

Procuramos, agora, nesta terceira parte de nosso trabalho, fazer uma aplicação prática do critério de classificação de questões — segundo seu poder discriminativo — proposto na segunda parte. Escolhemos, para isso, uma prova de Matemática, aplicada, entre outras, no exame vestibular da Escola Nacional de Ciências Estatísticas, realizado no início do ano de 1979.

A referida prova constou de vinte questões, cujos enunciados apresentamos a seguir:

N.º 1:

Calcule $\dfrac{1}{0{,}363636\ldots - 0{,}333\ldots}$

N.º 2:

Calcule a soma de todos os múltiplos de 6 que estejam entre 10^2 e 10^3.

N.º 3:

Determine k para que a equação $x^2 + kx + 1 = 0$ possua duas raízes reais distintas.

N.º 4:

Resolva a inequação

$$\frac{x + 1}{x - 1} \leqslant 1$$

N.º 5:

Determine as soluções reais do sistema de equações:

$$\begin{cases} x^2 = 2y \\ y^2 = 2x \end{cases}$$

N.º 6:

Determine a área da região do plano formada pelos pontos $(x; y)$ tais que $-2 \leqslant x \leqslant 4$ e
$-3 \leqslant y \leqslant 1$.

N.º 7:

A reta R passa pelos pontos $(-3; 1)$ e $(2; 2)$. A reta S é perpendicular a R e passa pelo ponto $(1; -2)$. Determine a equação da reta S.

N.º 8:

Determine o centro e o raio do círculo de equação

$$2x^2 + 2y^2 - 2x + 6y + 3 = 0$$

N.º 9:

Determine, em R^2, os vetores unitários paralelos à reta $y = 2x + 3$.

N.º 10:

Sendo $f(x) = x^2$, para $x \geqslant 0$, esboce o gráfico de $y = g(x)$, onde g é a função inversa de f.

N.º 11:

Determine o resto da divisão de $x^{80} + x^{40} - 2x + 1$ por $x^2 - 1$.

N.º 12:

Determine as três raízes complexas da equação $z^3 = -i$.

N.º 13:

Uma esfera de raio R é seccionada por um plano que dista x do seu centro. Calcule a área da seção obtida em função de R e x.

N.º 14:

Um trapézio retângulo tem bases que medem 4 e 9, e suas diagonais são perpendiculares. Determine sua altura.

N.º 15:

Um triângulo retângulo tem catetos que medem 3 e 4. Determine o cosseno do ângulo agudo que tem por lados a altura e a mediana relativas à hipotenusa.

N.º 16:

Num produto de dois fatores, o primeiro fator sofre um aumento de 20% e o segundo fator sofre uma diminuição de 20%. Determine a variação percentual sofrida pelo produto.

N.º 17:

No ano t, a população x de uma certa comunidade é dada pela fórmula:

$$x = \frac{2{,}4 \cdot 10^5}{1 + 3^{-0{,}02t}}$$

Calcule em que ano a população será de um milhão e oitocentos mil indivíduos.

N.º 18:

O técnico de uma equipe de futebol tem dois jogadores disponíveis para cada uma das onze posições de sua equipe. De quantos modos distintos pode ele escalar a sua equipe de onze jogadores?

N.º 19:

Colocando em ordem decrescente as soluções positivas da equação sen $\frac{1}{x} = 1$, determine a terceira solução.

N.º 20:

O traço da matriz $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ é definido como $\sum_{j=1}^{3} a_{jj}$. Calcule o traço do produto BC, onde $B = \begin{bmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 2 \end{bmatrix}$ e $C = \begin{bmatrix} 3 & 0 & 1 \\ -2 & 1 & -1 \end{bmatrix}$.

3.2 Critério de correção da prova

A cada uma das questões acima apresentadas foi atribuído o valor de 0,5 (meio) ponto.

Na correção dessas questões adotou-se o seguinte critério: às questões para as quais os candidatos apresentaram resposta certa e desenvolvimento certo foi atribuído o grau máximo: 0,5 ponto; às questões nas quais os candidatos cometeram alguma pequena falha quer na resposta quer no desenvolvimento foi atribuído o grau 0,4; às questões que os candidatos erraram, mas em cujo desenvolvimento encontrou-se alguma parte aproveitável, foi atribuído u mgrau variável entre 0,1 e 0,3; finalmente, às questões totalmente erradas ou não executadas receberam o grau 0 (zero).

3.3 Composição dos grupos

Do total de inscritos compareceram a essa prova 269 candidatos. Entretanto, apenas 218 deles obtiveram grau diferente de 0 (zero).

No pressuposto de que os que obtiveram grau 0 (zero) — ao todo 51 candidatos — não deviam estar em condições de realizar a prova, todos eles foram liminarmente excluídos do conjunto inicial. Também foram excluídos 18 candidatos que, embora tendo obtido grau diferente de zero, não conseguiram acertar, pelo menos, uma das vinte questões da prova.

Os candidatos restantes foram, em seguida, classificados na ordem decrescente do total de pontos obtido, tendo sido considerados "habilitados" os que lograram, pelo menos 30 (trinta) pontos. Em princípio, os resultados deveriam ter sido os seguintes:

Aprovados (com, pelo menos, 30 pontos): 77 candidatos;

Reprovados (com menos de 30 pontos): 123 candidatos.

A fim de acentuar a diferença entre os dois grupos, A e B, julgamos conveniente eliminar também todos os candidatos que alcançaram um total de pontos superior a 25 e inferior a 35. Dessa forma, ficaram excluídos 8 candidatos do grupo dos reprovados e 20 candidatos do grupo dos aprovados.

Os restantes candidatos aprovados e reprovados constituíram, respectivamente, os dois grupos A e B, conforme indicamos abaixo:

Grupo Superior, A: 57 candidatos, com grau mínimo 35.

Grupo Inferior, B: 115 candidatos, com grau máximo 25.

3.4 Análise e classificação das questões

Uma vez formados os dois grupos, procedemos à contagem, para cada questão, das freqüências de acertos N_{AC} e N_{BC}, sendo que, com esse objetivo, foram consideradas certas apenas as questões que receberam graus 0,5 e 0,4.

O quadro a seguir fornece um resumo dos resultados obtidos na análise das 20 questões da citada prova:

QUADRO 4

N.º DA QUESTÃO	FREQÜÊNCIAS DE ACERTOS		RAZÕES		VALOR DE r:	CLASSIFICAÇÃO DA QUESTÃO
	Grupo A N_{AC}	Grupo B N_{BC}	$R_1 = \frac{N_{AC}}{N_A}$	$R_2 = \frac{N_{BC}}{N_B}$	$r = \frac{R_1}{R_2}$	
1	41	48	0,72	0,42	1,71	DPf
2	39	20	0,68	0,17	4,00	DPM
3	40	10	0,70	0,09	7,78	DPF
4	38	9	0,67	0,08	8,40	DPF
5	37	14	0,65	0,12	5,42	DPM
6	44	32	0,77	0,28	2,75	DPM
7	46	28	0,81	0,24	3,38	DPM.
8	35	12	0,61	0,10	6,10	DPF
9	5	—	0,09	0,00	∞	DPF
10	29	12	0,51	0,10	5,10	DPM
11	14	1	0,25	0,01	25,00	DPF
12	17	2	0,30	0,02	15,00	DPF
13	20	5	0,35	0,04	8,75	DPF
14	18	5	0,32	0,04	8,00	DPF
15	31	3	0,54	0,03	18,00	DPF
16	26	21	0,46	0,18	2,56	DPM
17	32	11	0,56	0,10	5,60	DPM
18	27	16	0,47	0,14	3,36	DPM
19	12	—	0,21	0,00	∞	DPF
20	26	7	0,46	0,06	7,67	DPF

NOTA — $N_A = 57$ e $N_B = 115$

3.5 Conclusões

Todas as questões revelaram possuir poder discriminativo positivo. Apenas uma delas — a de n.º 1 — foi classificada como fracamente discriminativa positiva (DPf); oito questões — as de ns. 2, 5, 6, 7, 10, 16, 17 e 18 — foram classificadas como moderadamente discriminativas positivas (DPM); as onze restantes — de ns. 3, 4, 8, 9, 11, 12, 13, 14, 15, 19 e 20 — receberam a classificação de fortemente discriminativas positivas.

Sob o aspecto focalizado, pois, a prova que vimos de analisar cumpriu satisfatoriamente a sua finalidade.

A FUNÇÃO PERDA COMO FATOR NO TAMANHO DE UMA AMOSTRA

Ruy Donini Antunes *

SUMÁRIO

1. MODELO GERAL

1.1 Perda e Risco

Yates (4;292)[1] apresentou, pela primeira vez, uma solução generalizada do problema do dimensionamento de uma amostra com base na função perda. Daremos, a seguir, os detalhes de sua solução.

* Professor assistente do Departamento de Estatística do IME-USP.

[1] Do par de números colocados após o nome de algum autor, neste texto, o primeiro indica seu número de ordem na Bibliografia registrada ao fim do artigo, e o segundo indica a página de seu trabalho de onde foi aproveitado algum tópico para a presente redação. Alguns resultados mais importantes foram também numerados, porém sempre com um único número ao lado, entre parênteses.

Consideremos uma população **X** que dependa do parâmetro Θ, e seja $\hat{\Theta}$ um estimador justo desse parâmetro, definido num conjunto Ω. Uma genérica função perda será indicada por $L(\hat{\Theta};\Theta)$, e será definida por:

$$L(\hat{\Theta};\Theta) \begin{cases} = a_1(\hat{\Theta} - \Theta)^b \Longleftarrow \hat{\Theta} \geq \Theta \\ \\ = a_2(\Theta - \hat{\Theta})^b \Longleftarrow \hat{\Theta} < \Theta \end{cases}$$

onde a_1 a_2 e b são constantes (que dependem do problema onde a função perda está sendo aplicada), com a_1 e a_2 não negativos nem simultaneamente nulos, e $b > 0$.

A função perda traduz a perda em dinheiro (ou qualquer outra grandeza conversível em dinheiro) decorrente do erro cometido na estimação de Θ.

O risco associado ao estimador Θ e ao valor do parâmetro Θ será indicado por $R_L(\Theta)$, e é definido pela esperança matemática da função perda $L(\hat{\Theta};\ \Theta)$, com $\hat{\Theta}\ \varepsilon\ \Omega$. Isto é:

$$R_L(\Theta) = E\ [L(\hat{\Theta};\Theta)]$$

Admitiremos, no que se segue, que $\hat{\Theta}$ tenha distribuição normal, com função densidade representada por $f(\hat{\Theta};\ \Theta)$, $\hat{\Theta}\ \varepsilon\ \Omega$. Então:

$$R_L(\Theta) = \int_\Omega L(\hat{\Theta};\Theta)\ f(\hat{\Theta};\Theta)\ d\hat{\Theta} =$$

$$= \int_\Theta^{+\infty} a_1\ (\hat{\Theta} - \Theta)^b\ f(\hat{\Theta};\Theta)\ d\hat{\Theta} + \int_{-\infty}^{\Theta} a_2\ (\Theta - \hat{\Theta})^b\ f(\hat{\Theta};\Theta)\ d\hat{\Theta}$$

Fazendo a transformação $\hat{\Theta} - \Theta = = \sigma_o$, onde $g(z)$, $z\ \varepsilon\ R$, é a função densidade da variável normal padronizada Z, e $\sigma_o = + \sqrt{\sigma^2(\hat{\Theta})}$, resulta:

$$R_L(\Theta) = \sigma_o^b \left[a_1 \int_0^{+\infty} z^b\ g(z)\ dz + a_2 \int_{-\infty}^{0} (-z)^b\ g(z)\ dz \right]$$

Chamando a expressão entre colchetes de σ_o, obtemos então:

$$R_L(\Theta) = a_o\ \sigma_o^b$$

Para diferentes valores de b, o cálculo de

$$\int_0^{+\infty} z^b\ gz\ dz = \frac{1}{\sqrt{2\pi}} \int_0^{+\infty} z^b\ e^{-z^2/2}\ dz$$

que chamaremos de $I(b)$, depende da utilização da função gama:

$$\Gamma(m) = \int_0^{+\infty} x^{2m-1} e^{-x} dx, \quad m > 0$$

Mediante a transformação $x = z^2/2$, obtemos:

$$\Gamma(m) = \frac{1}{2^{m-1}} \int_0^{+\infty} z^{2m-1} e^{-z^2/2} dz$$

Fazendo $2m - 1 = b$, e introduzindo o fator $1/\sqrt{2\pi}$ em ambos os lados da igualdade, resulta:

$$I(b) = \frac{2^{\frac{b}{2}-1}}{\sqrt{\pi}} \Gamma\left(\frac{b+1}{2}\right) \tag{2}$$

O valor de $\tau\left(\frac{b+1}{2}\right)$ pode ser obtido em tabelas específicas; em caso de nenhuma estar disponível, damos os principais valores de $I(b)$. Os valores $I(1) = \frac{1}{\sqrt{2\pi}}$ e $I(2) = 1/2$ podem ser facilmente calculados; os demais foram calculados com base na tabela de Burington (2; 264).

TABELA 1

b	0,50	0,75	1,00	1,25	1,50	1,75	2,00
$I(b)$	0,4110	0,3986	0,3989	0,4097	0,4300	0,4599	0,5000

Na prática, $b = 1$ e $b = 2$ são os valores mais usados, pois, para tais valores, o parâmetro σ_o^b torna-se, respectivamente, o desvio padrão e a variância de $\hat{\Theta}$. Além do mais, para $b = 2$ ocorrem outras simplificações muito convenientes, que serão oportunamente citadas.

1.2 Dimensionamento da Amostra

Suponhamos que o custo do levantamento de uma amostra de tamanho n seja dado por uma função custo qualquer $C = C(n)$. Se não for considerada a presença do risco, isto é, se qualquer que seja a estimativa $\hat{\Theta}_2$ obtida com o estimador $\hat{\Theta}$, pode-se tomar como nulo o prejuízo decorrente do erro de estimação, o tamanho da amostra pode ser determinado mediante a simples fixação da verba C destinada à seleção.

No caso que nos interessa presentemente, entretanto, o custo de amostragem não será a única despesa associada à seleção da amostra;

devemos agregar-lhe o risco (ou seja, o prejuízo previsto) decorrente da utilização da estimativa $\hat{\Theta}_o$.

Seja então a função:

$$Q = R_L(\Theta) + C$$

onde poremos $R_L(\Theta) = R(C(n))$, isto é, o risco foi colocado como função direta do custo C de amostragem, e como função indireta (função de função) do tamanho n da amostra. Decorre que a quantia total empenhada nessa amostra de tamanho n será:

$$Q = Q(C) = Q(C(n)).$$

O problema consiste, basicamente, em minimizar Q.

Em caráter geral, esta minimização pode ser realizada em função de C, obtendo-se resultados genéricos que não exigem o conhecimento imediato da expressão formal da função custo C.

Sob este ponto de vista (minimização de Q em relação a C), Yates propõe para a variância do estimador Θ a forma genérica:

$$\sigma^2(\hat{\Theta}) = \sigma_o^2 = \frac{h}{C - c_o} - k \tag{3}$$

onde h e k são constantes que dependem da população amostrada e do método de amostragem escolhido, e c_o representa os custos fixos no processo de amostragem.

A função $Q(C)$ resultará então:

$$Q(C) = a_o \left(\frac{h}{C - c_o} - k \right)^{\frac{2}{b}} + C \tag{4}$$

e passará por um mínimo quando $\frac{dQ}{dC} = 0$, isto é, quando:

$$(C - c_o)_2 \left(\frac{h}{C - c_o} - k \right)^{1 \frac{2}{b}} = \frac{1}{2} a_o \; b \; h \tag{5}$$

Esta equação é uma equação em C; resolvida, dará o valor de C que minimiza a quantia empenhada na amostra (isto é, risco mais custo). Os cálculos de Q mínimo e de $R_L(\Theta)$ são imediatos.

É interessante observar que, se $b = 2$, as conclusões apresentadas podem ser obtidas mesmo dispensando a hipótese da normalidade da distribuição de $\hat{\Theta}$. De fato, para $b = 2$, suponhamos inicialmente uma função perda onde $a_1 = a_2$, e ambos iguais a um certo a:

$$L(\hat{\Theta}; \Theta) = a(\hat{\Theta} - \Theta)^2, \; \hat{\Theta} \in \Omega$$

O risco associado será:

$$R_L(\Theta) = E[L(\hat{\Theta};\Theta)] = a\ E[(\hat{\Theta} - \Theta)^2] = a\ \sigma^2(\hat{\Theta})$$

É, portanto, suficiente admitir que a distribuição de $\hat{\Theta}$ tenha variância finita.

Para uma função custo C, devemos minimizar $Q = R_L + C$. Tomando a forma genérica $\sigma^2(\Theta) = \dfrac{h}{C - c_o} - k$, o mínimo de Q ocorre quando:

$$(C - c_o)^2 = ah$$

A hipótese $a_1 = a_2$ é dispensável se a distribuição de $\hat{\Theta}$ for simétrica em relação a $E(\hat{\Theta}) = \Theta$. Obtemos:

$$R_L(\Theta) = \frac{1}{2}(a_1 + a_2) \cdot \sigma^2(\hat{\Theta})$$

e

$$(C - c_o)^2 = \frac{1}{2}(a_1 + a_2)\ h$$

Para a determinação do tamanho ótimo n da amostra, é necessário dispor da expressão de C em função de n, adotada como a mais conveniente para o caso. Se esta expressão estiver disponível desde o início, será mais adequado substituir C em (4), e calcular diretamente o valor de n, mediante a resolução da equação em n:

$$\frac{dQ}{dn} = 0$$

Observemos que todas as conclusões acima mencionadas são gerais, aplicando-se a qualquer processo de amostragem e a qualquer parâmetro Θ cuja variância do respectivo estimador varie em sentido contrário ao de n. Faremos a seguir algumas considerações mais detalhadas para a estimação da média μ de uma população, adotando amostragem casual simples e uma função custo linear, da forma $C = c_o + c_1\ n$, onde c_1 representa o custo unitário (constante) para amostrar uma unidade final (e inclui despesas do tipo pagamento de entrevistadores, impressão de questionários, tabulação, etc.).

2. AMOSTRAGEM CASUAL SIMPLES

2.1 Introdução

Já vimos que a função $Q(C)$, dada em (4), que representa a quantia total envolvida na seleção da amostra e na perda esperada pela tomada

de decisões em função dos resultados obtidos, passa por um mínimo quando ocorre (5).

No presente caso, utilizaremos o estimador

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

para estimar a média μ de uma população mediante a amostragem casual simples. Distinguiremos duas situações: amostragem com reposição (ou, equivalentemente, amostragem sem reposição em população infinita) e amostragem sem reposição (em população finita). Lembrando a expressão genérica da variância do estimador $\hat{\Theta}$, dada em (3), faremos agora as particularizações de h e k para $\hat{\Theta} = \overline{X}$, considerando as duas situações acima citadas.

2.2 Amostragem com Reposição (ou sem Reposição, em População Infinita)

Sabe-se que $\sigma^2(\overline{X}) = \frac{\sigma^2}{n}$ É conveniente então fazer:

$$h = \sigma^2 \, c_1 \quad \text{e} \quad k = 0$$

Os valores de b e a_o dependem da função perda adotada. Por exemplo, para $b = 2$ e $a_1 = a_2 = a$, resultará:

$$a_o = a\left[\int_0^{+\infty} z^2 \, g(z) \, dz + \int_{-\infty}^{0} z^2 \, g(z) \, dz\right]$$

isto é:

$$a_o = a\left[\frac{1}{2} + \frac{1}{2}\right]$$

ou seja:

$$a_o = a$$

Portanto, em (5), substituindo devidamente, obtemos:

$$(c_1 n)^2 \left(\frac{\sigma^2 c_1}{c_1 n} - 0\right)^{1\frac{2}{2}} = \frac{1}{2} \cdot a \cdot 2 \cdot \sigma^2 \cdot c_1$$

donde:

$$\boxed{n = \sqrt{\frac{a\,\sigma^2}{c_1}}} \qquad (6)$$

Se $a_1 \neq a_2$, então $a_o = \frac{1}{2}(a_1 + a_2)$, e portanto:

$$\boxed{n = \sqrt{\frac{(a_1 + a_2)\,\sigma^2}{2c_1}}}$$

Novamente, para $a_1 = a_2 = a$, e fazendo $b = 1$, a função perda será:

$$L(\bar{x};\ \mu) = a\,|\bar{x} - \mu|,\ \bar{x} \in \mathbf{R}$$

Então:

$$a_o = \frac{2a}{\sqrt{2\pi}}$$

e daí, substituindo em (5), teremos:

$$(c_1 n)^2 \left(\frac{\sigma^2}{n}\right)^{\frac{1}{2}} = \frac{1}{2} \cdot \frac{2a}{\sqrt{2\pi}} \cdot 1 \cdot \sigma^2\, c_1$$

isto é:

$$\boxed{n = \sqrt[3]{\left(\frac{a\,\sigma}{c_1\sqrt{2\pi}}\right)^2}} \tag{7}$$

Se $a_1 \neq a_2$, conclui-se que:

$$\boxed{n = \sqrt[3]{\left[\frac{(a_1 + a_2)\sigma}{2c_1\sqrt{2\pi}}\right]^2}} \tag{8}$$

2.3 Amostragem sem Reposição (em População Finita)

Nesta condição, sabe-se que

$$\sigma^2\overline{X} = \frac{\sigma^2}{n} \cdot \frac{N - n}{N - 1}$$

onde N é o tamanho da população. A fim de adaptar $\sigma^2(\overline{X})$ à expressão genérica em (3)), fica mais conveniente substituir σ^2 por $\frac{N-1}{N}S^2$, onde

$$S^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (X_i - \mu)^2$$

Resulta então:

$$\sigma^2(\overline{X}) = \frac{S^2}{n} \cdot \frac{N-n}{N} = \frac{S^2}{n} - \frac{S^2}{N}$$

Podemos então propor $h = S^2\, c_1$ e $k = \frac{S^2}{N}$.

Para $b = 2$ e $a_1 = a_2 = a$, substituindo em (5)), resultará:

$$(c_1 n)^2 = \frac{1}{2} \cdot a \cdot 2 \cdot S^2 c_1$$

donde:

$$\boxed{n = \sqrt{\frac{aS^2}{c_1}}} \tag{9}$$

Se, entretanto, $b = 1$ (ou qualquer outro valor diferente de 2), o valor de n não será tão facilmente encontrado. Ainda supondo, por questão de comodidade, que $a_1 = a_2 = a$, resulta para $b = 1$:

$$(c_1 n)^2 \left(\frac{S^2}{n} - \frac{S^2}{N}\right)^{\frac{1}{2}} = \frac{aS^2 c_1}{\sqrt{2\pi}}$$

Quadrando os dois membros, resultará a seguinte equação de 4.º grau em n:

$$n^3 - \frac{n^4}{N} - \frac{a^2 S^2}{2\pi c_1^2} \tag{10}$$

que poderá ser resolvida por aproximações sucessivas; o primeiro valor proposto a n poderá ser o valor calculado mediante a fórmula (7).

Analisando as fórmulas (6) e (9) (ou mesmo os resultados (7) e (10)), chegamos a uma curiosa conclusão: o tamanho da amostra é maior na amostragem sem reposição do que na amostragem com reposição, o que parece opor-se ao bom senso habitual. O que ocorre, em verdade, é que não é simplesmente a precisão do estimador $\overline{X}$ que está em jogo, mas a quantia Q (risco + custo): na amostragem sem reposição, apesar de o tamanho da amostra resultar maior, a quantia Q resulta menor (pois o aumento de n acarreta uma diminuição maior na variância do que o aumento do custo). Isto, além de ser uma vantagem em si mesma, traz a vantagem extra de a amostra ser muito mais precisa na estimação de μ.

2.4 Uma Aplicação

A presente aplicação foi proposta por Cochran (3;125) que, por sua vez, a adaptou de um artigo de Blythe (1;68).

O preço de venda de uma partida de toras de madeira é $U.T_x$, onde U é o preço por unidade de volume, e T_X o volume total das toras da partida. O volume de cada tora será representado pela variável X, com média μ e variância σ^2, ambas finitas e positivas.

O número N de toras da partida foi contado e o volume médio $\overline{X}$ por tora foi estimado através de uma amostra casual simples de n toras. A estimativa é feita pelo vendedor, sendo aceita provisoriamente pelo comprador. Posteriormente, o comprador deve verificar o volume exato da madeira comprada, e se houver pago mais do que aquilo que lhe tenha sido entregue, o vendedor se compromete a reembolsá-lo da diferença; se houver pago menos, o fato não será levado em consideração.

Estabeleçamos inicialmente a função perda do vendedor. O preço exato de venda da partida é $U.T_X = UN\,\mu$, enquanto que o preço estimado mediante a amostra é $UN\overline{x}$ (onde $N\overline{x}$ é uma estimativa de T_X). Então:

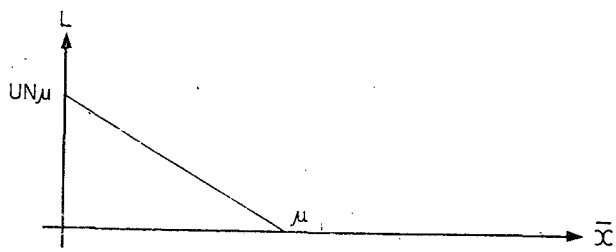
se $UN\overline{x} \geqslant UT_X$, o preço da partida foi superavaliado (ou avaliado corretamente); se foi superavaliado, o vendedor devolverá o excesso recebido, mas não perde nada;

se $UN\overline{x} < UT_X$, o preço da partida foi subavaliado, e o vendedor perderá a quantia $UT_X - UN\overline{x} = UN\,(\mu - \overline{x})$.

A função perda, que indicaremos por $L(\overline{x};\ \mu)$, será

$$L(\overline{x};\ \mu) \begin{cases} = 0 = 0 \cdot (\overline{x};\ -\mu) \Longleftarrow \overline{x} \geq \mu \\ = UN(\mu - \overline{x}) \Longleftarrow 0 < \overline{x} < \mu \end{cases}$$

Ou seja, uma função perda com $a_1 = 0$, $a_2 = UN$ e $b = 1$.



Para calcularmos agora o risco do vendedor (isto é, a perda prevista por transação), suporemos que $\bar{X}$ tenha distribuição normal.

Como já vimos em (1):

$$R_L(\Theta) = a_o \cdot \sigma_o^b$$

onde $a_o = (a_1 + a_2) \ . \ I(b)$, com $I(b)$ dado em (2).

No nosso caso, $a_1 = 0$ e $a_2 = UN$. Por outro lado, como $b = 1$, decorre que $I(b) = 1/\sqrt{2\pi}$ (conforme tabela 1), de modo que:

$$a_o = \frac{UN}{\sqrt{2\,\pi}}$$

Conforme a amostragem seja com ou sem reposição, teremos, respectivamente:

$$\sigma_o^b = \begin{cases} = \sigma(\overline{X}) = \dfrac{\sigma}{\sqrt{n}} \\ \text{ou} \\ = \sigma(\overline{X}) = S\sqrt{\dfrac{1}{n} - \dfrac{1}{N}} \end{cases}$$

e portanto o risco será, respectivamente:

$$R_L(\mu) \begin{cases} = \dfrac{UN\sigma}{\sqrt{2\,\pi\,n}} \\ \cdot \quad \text{ou} \\ = \dfrac{UNS}{\sqrt{2\,\pi\,n}}\sqrt{\dfrac{1}{n} - \dfrac{1}{N}} \end{cases}$$

A última providência consiste em dimensionar a amostra.

No primeiro caso (amostragem com reposição), conforme (8), vem:

$$\boxed{n = \sqrt[3]{\left(\frac{UN\sigma}{2c_1\sqrt{2\,\pi}}\right)^2}}$$

enquanto que no segundo caso (amostragem sem reposição), conforme (10), deveremos resolver a equação em n:

$$n^3 - \frac{n^4}{N} = \left(\frac{UNS}{2c_1\sqrt{2\,\pi}}\right)^2$$

BIBLIOGRAFIA

(1) BLYTHE Jr, R. H. (1945) — The Economics of Sample Size Applied to the Scaling of Sawlogs, in *Biometrics Bulletin,* vol. 1, n.º 6, pp. 67-70;

(2) BURINGTON, Richard Stevens (1957) — Handbook of Mathematical Tables and Formulas. Ed. Handbook Publishers, Inc., Sandusky, Ohio, EUA, 3.ª Edição;

(3) COCHRAN, William G. (1965) — Técnicas de Amostragem. Editora Fundo de Cultura, Brasil. Traduzido da 2.ª Ed. de *Sampling Techniques,* John Wiley & Sons, Inc., New York, 1963;

(4) YATES, Frank (1953) — Sampling Methods for Census and Surveys. Edição Hafner Publishing Company, New York, 2.ª Edição.

METODOLOGIA DO ÍNDICE NACIONAL DE PREÇOS AO CONSUMIDOR - INPC

RESOLUÇÃO — PR n.º 17, de 15-04-80 — Aprova metodologia de cálculo do INPC.

Aprova as metodologias de cálculo dos índices e de obtenção dos cadastros de produtos e de locais de compra, assim como os pesos utilizados para apuração do Índice Nacional de Preços ao Consumidor, a que se refere o Decreto n.º 84.560, de 14 de março de 1980.

1. CONSIDERAÇÕES INICIAIS

Em atendimento aos dispositivos legais expressos no Decreto n.º 84.560, de 14 de março de 1980, que regula a Lei 6.708, de 30 de outubro de 1979, que dispõe sobre a correção automática de salários, a Fundação Instituto Brasileiro de Geografia e Estatística — IBGE publica as informações básicas, a metodologia de cálculo e de obtenção dos cadastros de produtos e de locais, bem como os pesos utilizados no Índice Nacional de Preços ao Consumidor — INPC.

A escolha dos diversos métodos usados baseou-se na teoria e na prática dos Índices de Preços ao Consumidor.

1.1 Os Índices

O IBGE procede a cálculos de Índices de Preços ao Consumidor para cada uma das nove áreas metropolitanas brasileiras e Brasília e, com base nestes, elabora um Índice Nacional de Preços ao Consumidor.

* Transcrito do Boletim de Serviço n.º 1.444, IBGE.

1.2 Referência Geográfica

Todas as pesquisas utilizadas para o cálculo do INPC têm como referência geográfica a área metropolitana, isto é, as famílias selecionadas para os inquéritos básicos (Estudo Nacional da Despesa Familiar — ENDEF e Pesquisa de Locais de Compra — PLC) e os estabelecimentos constantes dos levantamentos contínuos (ESPECIFICAÇÃO e COLETA DE PREÇOS) pertencem à área urbana das nove áreas metropolitanas e Brasília, as quais compreendem um total de 116 municípios.

1.3 Referência Populacional

O INPC refere-se às famílias cuja principal fonte de rendimento é o trabalho assalariado e cujo total do rendimento monetário familiar disponível encontra-se entre um e cinco salários mínimos. Esta qualificação se manifesta principalmente nas estruturas de pesos utilizados nos cálculos dos índices para cada área, uma vez que são obtidas a partir dos domicílios que atendem às condições mencionadas.

1.4 Periodicidade

O INPC é uma estatística contínua, de periodicidade mensal para todas as áreas. Todos os produtos são pesquisados ao longo de cada mês, de modo a refletir o movimento de preços durante todo o período.

2. A METODOLOGIA DE CÁLCULO

2.1 Cálculos dos Índices de Preços ao Consumidor a Nível de Área

A metodologia de cálculo a nível de cada área pode ser dividida em três fases:

— em primeiro lugar para se obter o movimento de preços do subitem (menor desagregação dos orçamentos familiares) calculam-se os relativos de preços de todos os produtos/locais do subitem; obtêm-se as médias dos relativos a nível de produtos e depois a média de todos os produtos.

— em segundo lugar, os índices dos itens são calculados através de dois métodos: para a grande maioria dos itens (43) é utilizada a fórmula de Laspeyres, com pesos móveis, i.e.

$$I^{1}_{t,\,t-1} = \sum_{k=1}^{m} \omega^{ik}_{t-1} \cdot y^{ik}_{t,\,t-1}$$

onde ω_{t-}^{ik} é o peso do subitem k do item i

$y_{t,t-1}^{ik}$ é o relativo do subitem k do item i

$I_{t,t-1}^{i}$ é o índice do item i no mês "t" relativo ao mês $t-1$

$k = 1, \ldots m$ (m é o número de subitens do item i)

Para os itens sazonais — tubérculos, raízes e legumes; hortaliças e verduras e frutas — usa-se um painel de pesos mensais e aplica-se a fórmula de Paasche:

$$I_{t,t-1}^{i_s} = 1 \,/\, \sum_{k=1}^{m} \omega_t^{i_s k}\, y_{t,t-1}^{i_s k}$$

onde: $\omega_t^{i_s k}$ é o peso do subitem k do item sazonal i.

$y_{t,t-1}^{i_s k}$ é o relativo do subitem k do item sazonal i

$I_{t,t-1}^{i_s}$ é o índice do item sazonal i

— em terceiro lugar, o cálculo do índice para cada área é obtido pela aplicação da fórmula de Laspeyres (base de cálculo e base de ponderações móveis), isto é, por uma média ponderada dos índices dos itens, cujos pesos básicos são dados no anexo VI, ou seja:

$$I_{t,t-1}^{A} = \sum_{i=1}^{46} \omega_{t-1}^{i}\, I_{t,t-1}^{i}$$

onde ω_{t-1}^{i} é o paso do item i

$I_{t,t-1}^{i}$ é o índice do item i

$I_{t,t-1}^{A}$ é o índice de Preços ao Consumidor da área A.

$i = 1, \ldots, 46$

2.2 Cálculo do Índice Nacional de Preços ao Consumidor — INPC

Uma vez obtidos os índices relativos a cada área metropolitana, o INPC é calculado pela média aritmética ponderada dos índices regionais, sendo os pesos dados pela população residente em 1975, estimada pelo IBGE, conforme o anexo I. Analiticamente:

$$INPC_{t,t-1} \sum_{A=1}^{10} I_{t,t-1}^{A}\, v^{A}$$

onde INPC $t, t-1$ = índice Nacional de Preços ao Consumidor do mês t relativo ao mês $t-1$.

$I^{A\ 1}_{t,t-1}$ — índice de Preços ao Consumidor Restrito da área metropolitana "A" do mês t relativo ao mês $t-1$.

Peso da área metropolitana "A"

$A = 1, \ldots, 10$

3. OBTENÇÃO DOS CADASTROS DE PRODUTOS E LOCAIS

3.1 Pesquisa de Especificação de Produtos: Cadastro de Produtos

Para se obter a estimação do movimento de preços é necessária a especificação dos bens e serviços constantes de uma cesta de mercadorias, a fim de que se possa acompanhar os mesmos produtos ao longo do tempo. Então um trabalho de campo é feito tomando como guia essencial os próprios agregados da ENDEF (subitens).

A apuração destas informações dá origem ao CADASTRO DE PRODUTOS da respectiva área metropolitana, obtido após a seleção de determinado número de produtos bem especificados para cada subitem do índice. A dimensão e o conteúdo deste cadastro variam conforme a área metropolitana. Em média, temos em torno de 2.000 (dois mil) produtos exaustivamente descritos em seus atributos determinantes de preços.

3.2 Pesquisa de Locais de Compra: Cadastro de Informantes

A fim de se montar uma coleta sistemática de preços é necessário uma amostra de locais obtida por métodos estatísticos rigorosos. A Pesquisa de Locais de Compra — PLC é o instrumento utilizado para se obter o "universo" de informantes a partir do qual se extrai a amostra de locais.

Com base na Pesquisa Nacional por Amostragem de Domicílios (FNAD) definiram-se duas subamostras para a aplicação de PLC, cujo tamanho para dez áreas é dado no anexo II.

Às famílias selecionadas indaga-se o endereço onde adquirem as diversas categorias de bens e serviços. O resultado é um cadastro de informantes (cujos resultados agregados encontram-se no anexo III) para cada item de consumo, onde estão associadas, a cada estabelecimento, as freqüências com que foram apontados pelas famílias.

A partir deste esquema (categorias de consumo/local/freqüência) seleciona-se, com probabilidade determinada pela freqüência relativa, a amostra de estabelecimento para cada categoria de consumo, cujos resultados agregados estão no anexo IV.

3.3 A Pesquisa Contínua de Preços

A última fase de implantação do sistema constitui-se da pesquisa contínua de preços, de ciclo mensal, ao longo do mês, e que gera a série histórica de preços ao consumidor.

Os instrumentos de campo são os questionários de coleta de preços ao consumidor gerados a partir da fusão dos cadastros de locais e produtos e emitidos por computador. Paralelamente, é posto em prática em cada trimestre um esquema de ATUALIZAÇÃO que permite a absorção imediata das alterações de forma e conteúdo dos produtos (contemplando inclusive novos produtos), bem como a eventual substituição de locais.

4. DETERMINAÇÃO DOS PESOS: ESTUDO NACIONAL DA DESPESA FAMILIAR — ENDEF

O Estudo Nacional da Despesa Familiar — ENDEF é uma pesquisa de orçamentos familiares de caráter nacional aplicada no período de 19 de agosto de 1974 a 11 de agosto de 1975. Sua amostra foi desenhada de tal modo que os resultados fossem representativos a nível de área metropolitana. No total das áreas abrangidas pelo INPC foram pesquisados 22.788 domicílios, distribuídos conforme o anexo V.

As informações dos domicílios que atendem à definição do INPC — chefes assalariados em sua ocupação principal com rendimentos compreendidos entre 1 e 5 salários mínimos — são utilizadas para a montagem, em cada área, das estruturas de pesos, cujos dados a nível de item apresentam-se no anexo VI. Observe-se que o agregado "item" compõe-se de um conjunto de subitens, variando sua composição de acordo com a área.

5. ÍNDICES NACIONAIS DE PREÇOS AO CONSUMIDOR SEMESTRAIS

Para atender ao artigo 2.º, item III, § 1.º da Lei 6.708, estima-se os índices semestrais de preços ao consumidor pela acumulação geométrica dos índices mensais definidos conforme o item 2.2 desta metodologia. Os resultados são divulgados no Diário Oficial da União, por resolução do Presidente do IBGE.

ANEXO I

POPULAÇÃO POR ÁREA METROPOLITANA — 1975

ÁREA METROPOLITANA	POP	%
Belém	800.482	0,0270
Fortaleza	1.317.496	0,0444
Recife	2.153.435	0,0726
Salvador	1.401.228	0,0472
Belo Horizonte	2.022.846	0,0682
Rio de Janeiro	8.328.784	0,2806
São Paulo	10.041.132	0,3383
Curitiba	1.013.279	0,0341
Porto Alegre	1.836.179	0,0619
Brasília (DF)	763.254	0,0257
Σ	29.678.115	1

FONTE — Anuário Estatístico do Brasil — 1975, IBGE.
ELABORAÇÃO — Departamento de Estatísticas e Índices de Preços — DESIP, IBGE.

ANEXO II

NÚMERO DE DOMICÍLIOS VISITADOS NA PESQUISA DE LOCAIS DE COMPRA — PLC, POR ÁREA METROPOLITANA

ÁREA METROPOLITANA	NÚMERO DE DOMICÍLIOS DA PLC
Belém	734
Fortaleza	633
Recife	1.541
Salvador	877
Belo Horizonte	1.560
Rio de Janeiro	3.885
São Paulo	2.666
Curitiba	746
Porto Alegre	1.529
Brasília	898

FONTE — Pesquisa de Locais de Compra — PLC, Departamento de Estatísticas e Índices de Preços — DESIP, IBGE.
ELABORAÇÃO — Departamento de Estatísticas e Índices de Preços — DESIP, IBGE.

ANEXO III

NÚMERO TOTAL DE LOCAIS DE COMPRA OBTIDO NA PESQUISA DE LOCAIS DE COMPRA — PLC, POR ÁREA METROPOLITANA

ÁREA METROPOLITANA	NÚMERO DE LOCAIS DE COMPRA
Belém	2.406
Fortaleza	3.284
Recife	5.329
Salvador	3.632
Belo Horizonte	7.152
Rio de Janeiro	7.092
São Paulo	12.466
Curitiba	3.180
Porto Alegre	7.519
Brasília	3.134
TOTAL	55.194

FONTE — Pesquisa de Locais de Compra — PLC — Departamento de Estatísticas e Índices de Preços — DESIP, IBGE.

ELABORAÇÃO — Departamento de Estatísticas e Índices de Preços — DESIP, IBGE.

ANEXO IV

NÚMERO TOTAL DE LOCAIS DA AMOSTRA EXTRAÍDOS DA PESQUISA DE LOCAIS DE COMPRA E EXTRA — PLC, POR ÁREA METROPOLITANA

ÁREA METROPOLITANA	NÚMERO DE LOCAIS		TOTAL
	PLC	Extra-PLC (1)	
Belém	686	15	701
Fortaleza	1.195	15	1.210
Recife	994	76	1.070
Salvador	1.084	15	1.099
Belo Horizonte	1.322	36	1.358
Rio de Janeiro	1.538	186	1.724
São Paulo	1.883	31	1.914
Curitiba	1.093	15	1.108
Porto Alegre	1.315	54	1.369
Brasília	968	14	982
Σ	12.078	457	12.535

FONTE — Pesquisa de Locais de Compra — PLC, 1978/79/80 — Departamento de Estatísticas e Índices de Preços — DESIP, IBGE.

ELABORAÇÃO — Departamento de Estatísticas e Índices de Preços — DESIP, IBGE.

NOTA — Estas amostras constituem o primeiro painel. Ao longo do tempo foram feitos acréscimos de locais para itens específicos.

(1) Obtidos de outras fontes que não a PLC.

ANEXO V

NÚMERO DE DOMICÍLIOS POR ÁREA METROPOLITANA

ÁREA METROPOLITANA	NÚMERO DE DOMICÍLIOS PESQUISADOS
Rio de Janeiro	3.496
São Paulo	2.864
Curitiba	1.863
Porto Alegre	1.973
Belo Horizonte	1.892
Fortaleza	1.896
Recife	2.143
Salvador	2.088
Belém	1.907
Brasília (DF)	2.566
TOTAL	22.788

FONTE — Estudo Nacional de Despesa Familiar, Fundação IBGE, 1974/1975.
ELABORAÇÃO — *Departamento de Estatísticas e Índices de Preços* — DESIP, IBGE.

ANEXO VI

ESTRUTURAS BÁSICAS DE PONDERAÇÕES

ITENS	ÁREAS METROPOLITANAS — PONDERAÇÕES (%)									
	Rio de Janeiro	Porto Alegre	Belo Horizonte	Recife	São Paulo	Brasília	Belém	Fortaleza (1)	Salvador (1)	Curitiba (1)
Cereais, Leguminosas e Oleaginosas	6,10	4,82	7,46	4,31	6,13	8,06	4,75	5,02	2,55	3,79
Farinhas, Féculas e Massas	1,46	1,51	1,80	4,77	0,99	1,16	7,62	2,05	2,60	2,57
Tubérculos, Raízes e Legumes	2,41	1,94	2,45	2,48	1,74	2,02	1,28	2,81	1,92	2,33
Açúcar e Derivados	1,66	2,15	2,24	2,42	1,22	1,77	1,68	4,96	1,69	2,04
Hortaliças e Verduras	0,38	0,37	0,41	0,38	0,53	0,28	0,29	0,04	0,18	0,54
Frutas	1,48	0,90	1,06	1,58	1,60	1,35	1,24	3,63	1,96	2,55
Carnes Frescas e Vísceras	5,70	7,23	4,51	5,20	5,40	6,48	11,12	14,52	22,75	18,34
Pescado	0,95	0,20	0,27	0,97	0,43	0,38	3,64	2,13	1,87	0,03
Carnes e Peixes Industrializados	1,75	1,27	2,04	3,83	1,72	0,58	3,42	0,24	3,07	0,16
Aves e Ovos	3,09	2,82	2,82	4,50	2,67	2,36	2,47	2,10	1,71	2,18
Leite e Derivados	3,75	4,58	4,23	4,94	4,09	4,32	4,04	4,59	3,39	3,95
Panificados	3,57	3,25	3,24	7,47	3,23	3,22	4,30	3,69	3,10	2,45
Óleo e Gorduras	2,06	2,24	2,25	0,89	1,84	2,14	1,12	1,08	0,69	1,60
Bebidas não Alcoólicas e Infusões	2,03	2,04	1,89	2,36	1,65	1,78	2,84	4,29	4,44	4,72
Enlatados e Conservas	0,33	0,31	0,33	0,43	0,40	0,19	0,35	0,25	0,42	0,58
Sal e Condimentos	0,74	0,57	0,80	0,67	0,65	0,61	0,48	0,79	0,66	1,13
Alimentação Fora do Domicílio	5,71	4,53	3,96	3,69	4,41	5,99	2,42	3,80	0,60	0,35
Habitação	7,85	6,13	6,73	5,47	11,20	5,00	5,01	9,64	11,35	9,18
Reparos	1,19	1,70	3,05	1,39	1,64	2,79	1,14	—	—	—
Artigos de Limpeza	2,50	2,49	2,61	2,65	2,81	3,02	2,43	0,69	1,37	1,63
Combustíveis	0,66	1,85	0,61	1,26	1,13	0,93	1,10	0,11	1,26	0,39
Serviços Públicos	5,21	4,28	5,32	4,11	5,00	4,82	5,04	6,62	4,57	4,42
Mobiliário	1,55	1,79	1,65	1,91	2,21	2,37	1,57	2,21	2,26	1,80
Utensílios e Enfeites	0,45	0,50	0,47	0,42	0,40	0,47	0,43	0,39	0,42	0,39
Artigos de Cama, Mesa e Banho	0,67	0,96	0,99	0,62	0,81	1,01	0,46	0,16	0,38	0,42
Eletrodomésticos e Equipamentos	1,17	1,98	1,80	1,39	1,73	1,84	1,74	1,43	0,97	1,63
TV e Som	1,88	1,65	1,97	1,52	2,06	2,90	1,78	0,10	0,89	1,13
Roupas de Homem	1,88	1,95	1,66	1,77	1,67	1,99	1,75	2,11	1,88	1,62
Roupas de Mulher	1,55	1,68	1,11	1,00	1,63	1,04	1,07	1,14	1,57	1,43
Roupas de Criança	0,95	1,57	1,32	1,11	1,19	1,62	1,05	0,53	0,64	0,63
Calçados e Outros Apetrechos	1,65	1,77	1,76	1,65	1,75	1,65	1,55	1,48	1,85	1,13
Joias e Bijuterias	1,08	1,17	1,08	0,71	1,14	1,32	0,82	—	—	—
Tecidos e Artigos de Armarinho	0,66	0,95	1,37	1,11	0,46	1,02	1,06	0,47	0,51	0,23
Transporte Público	7,58	6,00	6,01	6,06	5,35	6,28	5,84	2,36	3,31	3,74
Veículo Próprio	0,91	2,49	1,63	0,51	3,37	1,64	0,21	2,03	0,47	4,57
Comunicações	0,20	0,02	0,14	0,06	0,11	0,21	0,09	0,02	0,02	0,03
Produtos Farmacêuticos	2,24	2,74	2,64	2,47	2,53	2,00	1,81	0,86	1,35	1,48
Óculos e Lentes	0,21	0,12	0,27	0,09	0,13	0,12	0,03	0,27	—	0,97
Atendimento Médico	0,55	0,61	0,62	0,17	0,77	0,48	0,16	0,35	0,59	1,60
Serviços Médicos	0,27	0,15	0,27	0,10	0,23	0,16	0,13	0,02	—	—
Higiene Pessoal	2,89	3,25	2,38	3,35	2,92	3,09	3,45	2,00	3,24	3,79
Serviços Pessoais	1,84	2,22	2,75	1,63	1,82	2,12	1,18	9,07	2,32	2,99
Recreação	0,94	1,19	1,39	0,99	0,69	0,92	1,19	0,61	0,66	0,61
Fumo e Álcool	5,92	6,32	5,03	4,38	5,24	5,08	3,59	4,29	3,15	0,63
Educação	1,49	1,58	1,96	1,06	1,09	1,33	1,07	2,77	1,08	1,38
Leitura e Papelaria	0,29	0,18	0,17	0,15	0,22	0,11	0,19	0,08	0,23	0,21

FONTE — IBGE.
ELABORAÇÃO — Departamento de Estatísticas e Índices de Preços — DESIP, IBGE.
(1) Estruturas baseadas em pesquisa realizada pelo Ministério do Trabalho. A partir de agosto de 1980 serão utilizadas as novas ponderações do ENDEF.

Bibliografia

PUBLICAÇÕES DE INTERESSE PARA A ESTATÍSTICA EDITADAS POR ÓRGÃOS DO IBGE NO PERÍODO DE OUTUBRO DE 1979 A MARÇO DE 1980 *

31(81)(05)

BOLETIM ESTATÍSTICO. Rio de Janeiro, v. 36, n. 141/144, jan./dez. 1978.

31:325.11(81-24)

O quadro das famílias em domicílios de chefe migrante e natural; um estudo censitário dos diferenciais nas regiões metropolitanas brasileiras. Rio de Janeiro, 1979. 149 p., gráf., tab. (estudos e pesquisas, 2).

31:336.12(81)(058)

ESTATÍSTICAS ECONÔMICAS DO GOVERNO ESTADUAL E MUNICIPAL, Rio de Janeiro, v. 1, 1975, t. 1: Balanços Estaduais e Municipais.

31:338.3:63(81)(05)

LEVANTAMENTO SISTEMÁTICO DA PRODUÇÃO AGRÍCOLA; pesquisa mensal de previsão e acompanhamento das safras agrícolas, Rio de Janeiro, jul. 1979.

———. ago. 1979.

———. set. 1979.

———. out. 1979.

———. nov. 1979.

———. dez. 1979.

———. jan. 1980.

31:338.3:63(811)(058)

PRODUÇÃO AGRÍCOLA MUNICIPAL; culturas temporárias e permanentes, Rio de Janeiro, v.

* Preparado na Divisão de Informações Correntes do Departamento de Informação da Biblioteca Central do IBGE pela bibliotecária Isis Soares da Silva.

5, 1978, t. 1: Rondônia, Acre, Amazonas, Roraima, Pará, Amapá.

31:338.3:63(813.4/814.2) (058)

———. t. 3: Pernambuco, Alagoas, Sergipe, Bahia.

31:338.3:630.8(81) (058)

PRODUÇÃO EXTRATIVA VEGETAL; Brasil, Rio de Janeiro, v. 3, 1975.

———. v. 4, 1976.

———. v. 5, 1977.

31:63(81) (05)

NOTAS SOBRE OS PRINCIPAIS ACONTECIMENTOS NA AGRICULTURA BRASILEIRA, Rio de Janeiro, v. 5-6, jul. — out. 1979; v. 7, jan. 1980.

31:654.15(81) (058)

EMPRESAS TELEFÔNICAS, Rio de Janeiro, v. 8, 1977.

311.213.1:62/69
(811.1+811.4+811.6)

Censo industrial: *Rondônia, Roraima, Amapá*. Rio de Janeiro, 1979. 371 p., tab. (censos econômicos 1975: série regional, v. 2, t. 1).

311.213.1:62/69(811.2)

———: *Acre*. Rio de Janeiro, 1979. 173 p., tab. (censos econômicos 1975: série regional, v. 2, t. 2).

311.213.1:62/69(811.3)

———: *Amazonas*. Rio de Janeiro, 1979. 181 p., tab. (censos econômicos 1975: série regional, v. 2, t. 3).

311.213.1:62/69(811.5)

———: *Pará*. Rio de Janeiro, 1979. 191 p., tab. (censos econômicos 1975: série regional, v. 2, t. 4).

311.213.1:62/69(812.1)

———: *Maranhão*. Rio de Janeiro, 1979. 197 p., tab. (censos econômicos 1975: série regional, v. 2, t. 5).

311.213.1:62/69(812.2)

———: *Piauí*. Rio de Janeiro, 1979. 195 p., tab. (censos econômicos 1975: série regional, v. 2, t. 6).

311.213.1:62/69(813.1)

———: *Ceará*. Rio de Janeiro, 1979. 205 p., tab. (censos econômicos 1975: série regional, v. 2, t. 7).

311.213.1:62/69(813.2)

———: *Rio Grande do Norte*. Rio de Janeiro, 1979. 199 p., tab. (censos econômicos 1975: série regional, v. 2, t. 8).

311.213.1:62/69(813.3)

———: *Paraíba*. Rio de Janeiro, 1979. 203 p., tab. (censos econômicos 1975: série regional, v. 2, t. 9).

311.213.1:62/69(813.4)

———: *Pernambuco*. Rio de Janeiro, 1979. 217 p., tab. (censos econômicos 1975: série regional, v. 2, t. 10).

311.213.1:62/69(813.5)

———: *Alagoas*. Rio de Janeiro, 1979. 191 p., tab. (censos econô-

micos 1975: série regional, v. 2, t. 11).

311.213.1:62/69(814.1)

———: *Sergipe.* Rio de Janeiro, 1979. 189 p., tab. (censos econômicos 1975: série regional, v. 2, t. 12).

311.213.1:62/(814.2)

———: *Bahia.* Rio de Janeiro, 1979. 247 p., tab. (censos econômicos 1975: série regional, v. 2, t. 13).

311.213.1:62/69(816.2)

———: *Paraná.* Rio de Janeiro, 1979. 257 p., tab. (censos econômicos 1975: série regional, v. 2, t. 18).

311.213.1:62/69(816.4)

———: *Santa Catarina.* Rio de Janeiro, 1979. 239 p., tab. (censos econômicos 1975: série regional, v. 2, t. 19).

311.213.1:62/69(817.1)

———: *Mato Grosso do Sul.* Rio de Janeiro, 1979. 185 p., tab. (censos econômicos 1975: série regional, v. 2, t. 21).

311.213.1:62/69(817.2)

———: *Mato Grosso.* Rio de Janeiro, 1979. 181 p., tab. (censos econômicos 1975: série regional, v. 2, t. 22).

311.213.1:62/69(817.3)

———: *Goiás.* Rio de Janeiro, 1979. 223 p., tab. (censos econômicos 1975: série regional, v. 2, t. 23).

311.213.1:62/69(817.4)

———: *Distrito Federal.* Rio de Janeiro, 1979. 173 p., tab. (censos econômicos 1975: série regional, v. 2, t. 24).

311.213.1:63(81)

Censo agropecuário: Brasil. Rio de Janeiro, 1979. 471 p., tab. (censos econômicos 1975: série nacional, v. 1).

311.213.1:63
(811.1+811.4+811.6)

———: *Rondônia, Roraima, Amapá.* Rio de Janeiro, 1979. 535 p., tab. (censos econômicos 1975: série regional, v. 1, t. 1).

311.213.1:63(811.2)

———: *Acre.* Rio de Janeiro, 1979. 178 p., tab. (censos econômicos 1975: série regional, v. 1, t. 2).

311.213.1:63(811.3)

———: *Amazonas.* Rio de Janeiro, 1979. 31 p., tab. (censos econômicos 1975: série regional, v. 1, t. 3).

311.213.1:63(811.5)

———: *Pará.* Rio de Janeiro, 1979. 484 p., tab. (censos econômicos 1975: série regional, v. 1, t. 4).

311.213.1:63(812.1)

———: *Maranhão.* Rio de Janeiro, 1979. 502 p., tab. (censos econômicos 1975: série regional, v. 1, t. 5).

311.213.1:63(812.2)

———: *Piauí.* Rio de Janeiro, 1979. 520 p., tab. (censos econômicos 1975: série regional, v. 1, t. 6).

311.213.1:63(813.1)

———: *Ceará*. Rio de Janeiro, 1979. 696 p., tab. (censos econômicos 1975: série regional, v. 1, t. 7).

311.213.1:63(813.2)

———: *Rio Grande do Norte*. Rio de Janeiro, 1979. 504 p., tab. (censos econômicos 1975: série regional, v. 1, t. 8).

311.213.1:63(813.3)

———: *Paraíba*. Rio de Janeiro, 1979. 651 p., tab. (censos econômicos 1975: série regional, v. 1, t. 9).

311.213.1:63(813.4)

———: *Pernambuco*. Rio de Janeiro, 1979. 669 p., tab. (censos econômicos 1975: série regional, v. 1, t. 10).

311.213.1:63(814.2)

———: *Bahia*. Rio de Janeiro, 1979. 2 v., tab. (censos econômicos 1975: série regional, v. 1, t. 13 pt. 1-2).

311.213.1:63(815.1)

———: *Minas Gerais*. Rio de Janeiro, 1979. 2 v., tab. (censos econômicos 1975: série regional, v. 1, t. 14, pt. 1-2).

311.213.1:63(815.2)

———: *Espírito Santo*. Rio de Janeiro, 1979. 381 p., tab. (censos econômicos 1975: série regional, v. 1, t. 15).

311.213.1:63(815.3)

———: *Rio de Janeiro*. Rio de Janeiro, 1979. 343 p., tab. (censos econômicos 1975: série regional, v. 1, t. 16).

311.213.1:63(816.1)

———: *São Paulo*. Rio de Janeiro, 1979. 2 v., tab. (censos econômicos 1975: série regional, v. 1, t. 17, pt. 1-2).

311.213.1:63(816.2)

———: *Paraná*. Rio de Janeiro, 1979. 2 v., tab. (censos econômicos 1975: série regional, v. 1, t. 18, pt. 1-2).

311.213.1:63(816.5)

———: *Rio Grande do Sul*. Rio de Janeiro, 1979. 920 p., tab. (censos econômicos 1975: série regional, v. 1, t. 20).

311.213.1:63(817.2)

———: *Mato Grosso*. Rio de Janeiro, 1979. 257 p., tab. (censos econômicos 1975: série regional, v. 1, t. 22).

311.213.1:63(817.3)

———: *Goiás*. Rio de Janeiro, 1979. 704 p., tab. (censos econômicos 1975: série regional, v. 1, t. 23).

311.213.1:63(817.4)

———: *Distrito Federal*. Rio de Janeiro, 1979. 129 p., tab. (censos econômicos 1975: série regional, v. 1, t. 24).

312(81)(05)

BOLETIM DEMOGRÁFICO, Rio de Janeiro, v. 9, n. 3-4, jan./mar. — abr./jun. 1979; v. 10, n. 1-2, jul./set. — out./dez. 1979.

Composto e impresso no
Centro de Serviços Gráficos
do IBGE, Rio de Janeiro, RJ

IBGE

Presidente: Jessé Montello

Diretor-Técnico: Marco Antonio de Souza Aguiar
Diretor de Geodésia e Cartografia: Mauro Pereira de Mello
Diretor de Administração: Horácio de Almeida Amaral
Diretor de Formação e Aperfeiçoamento de Pessoal: Getúlio Pereira Carvalho
Diretor de Informática: Nelson Hochman
Diretor de Divulgação: Paulo Roberto Salema Garção Ribeiro