REVISTA BRASILEIRA DE ESTATÍSTICA

MINISTÉRIO DA ECONOMIA, FAZENDA E PLANEJAMENTO
FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 0034-7175

R. bras. Estat. Rio de Janeiro, v.50, nº 193, p. 1-129, jan./jun. 1989.

REVISTA BRASILEIRA DE ESTATÍSTICA

Órgão oficial do IBGE e da Sociedade Brasileira de Estatística

Publicação semestral, editada pelo IBGE, que se destina a promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais, através de divulgação de artigos inéditos.
Temas, abordando aspectos do desenvolvimento metodológico, serão aceitos desde que relevantes para os órgãos produtores de estatísticas.
Os originais para publicação deverão ser submetidos em 3 vias (que não serão devolvidas) para:

Djalma G. C. Pessoa
Editor Responsável - RBEs
ENCE
Rua André Cavalcanti, 106 - Bairro de Fátima
CEP 20 231 - Rio de Janeiro - RJ
- Os artigos submetidos à RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.
- Cada autor receberá, gratuitamente, 20 separatas de seu artigo.

Pedidos de assinatura, número avulso ou atrasado e para maiores informações dirigir-se ao
Centro de Documentação e Disseminação de Informações,
Rua General Canabarro, 666 - Bloco A - 2º andar - Maracanã
Tel.: (021)284-7690 e 228-9575 - CEP 20271 - Rio de Janeiro, RJ - Brasil

Editorada pelo CDDI - Departamento de Editoração
em novembro de 1990.

Capa
Pedro Paulo Machado

Revista brasileira de estatística / Fundação Instituto Brasileiro de Geografia e Estatística - ano 1, n.1(jan./mar., 1940)-
Rio de Janeiro: IBGE, 1940-
Trimestral. (1940-1986), semestral (1987-)
Órgão oficial do IBGE e da Sociedade Brasileira de Estatística.
Continuação de: Revista de economia e estatística
Sumários e índices acumulados de autor - assunto: v. 1/40 (1940) nov. 43, n. 169 (jan./mar. 1982)
ISSN 0034-7175 - Revista brasileira de estatística
1.Estatística - Periódicos. I. IBGE

IBGE. Gerência de Documentação e Biblioteca
RJ-IBGE/88-05

CDU 31(05)

Impresso no Brasil/Printed in Brazil

SUMÁRIO

NOTA DO EDITOR

A Revista Brasileira de Estatística completa, neste número, cinqüenta anos de existência. Ao longo desse tempo tem prestado inestimável serviço à comunidade científica do país. Grandes vultos do nosso meio científico divulgaram suas idéias através das páginas deste veículo. Ao longo de sua história tem-se concentrado, principalmente, nas seguintes áreas: Demografia; Metodologia Estatística, incluindo Teoria da Probabilidade e Matemática; Modelos e Análise Estatística em Ciências e, particularmente, em Economia.

Por ser a RBEs mantida pelo IBGE, grande parte da matéria publicada é dedicada à produção de estatísticas no país. Ela foi responsável, inclusive, pela divulgação de técnicas de amostragem na produção de dados, quando isto era novidade até no exterior.

Ao longo deste meio século, a RBEs, dentro de sua linha editorial, privilegiou muitos temas emergentes de interesse da Administração Pública e outra não é sua postura atual, que continua receptiva às colaborações voltadas para a solução de questões significantes para a Sociedade.

A Revista Brasileira de Estatística é, na espécie, a mais antiga em circulação na América Latina e sua longa existência é motivo de orgulho e de júbilo para toda a comunidade científica.

No próximo número, comemorativo dos 50 anos, a RBEs publicará, em fac-simile, artigos das diversas áreas de concentração da Revista, selecionados entre os mais expressivos.

OUTLIERS E ROBUSTEZ*

Oscar H. Bustos**

9 - TESTES DE DISCREPÂNCIA: CRITÉRIOS E IDÉIAS BÁSICAS

Obviamente, antes de decidir o que fazer com os *outliers* é necessário aplicar à amostra algum critério objetivo que permita assegurar se existem e quais são os pontos *outliers*. Aqui vamos entender por tais as observações "contaminantes", quer dizer, aquelas que provêm de uma distribuição diferente da hipotética (F_H), que postulamos se ajustar à maioria dos dados. Esses critérios objetivos ou procedimentos de detecção são chamados modernamente "testes de discordância ou discrepância" (Barnett e Lewis (1984)). Eles são freqüentemente denominados "testes de rejeição" de *outliers*, mas esta nomenclatura não é de todo adequada, pois nem sempre é rejeitado o (ou os) *outlier* detectado.

Para fixar idéias seguiremos trabalhando com o modelo proposto no começo da Seção 8. Vamos supor agora que as observações $x_1, \ldots, x_n$ são realizações de n variáveis aleatórias independentes $X_1, \cdots, X_n$, mas que poderiam não ter uma distribuição comum. É claro que, para construir um teste com propriedades "matematicamente" demonstráveis, necessitamos especificar melhor esta hipótese alternativa. Porém, mesmo sem especificá-la, algo pode ser dito. Com efeito, em muitas situações

*Nota do editor. Escrito a convite, este artigo objetiva oferecer uma retrospectiva sobre "Outliers e Robustez". Dividido em 3 partes, a 1ª foi publicada no número 191, a 2ª no número 192 e, a 3ª e última nesta edição.

**Pesquisador do Instituto de Matemática Pura e Aplicada - IMPA.

é possível sugerir um teste de discrepância razoável (outra coisa será estabelecer suas propriedades de um ponto de vista formal).

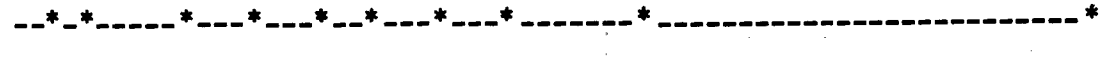
Suponhamos que $n = 10$ e que as observações são:

790, 750, 910, 650, 990, 630, 1290, 820, 860, 710.

Sempre será conveniente fazer alguns gráficos para ajudar-nos na escolha de um procedimento adequado. As observações ordenadas, denotadas por $x_{(1)}, \cdots, x_{(10)}$, são:

630, 650, 710, 750, 790, 820, 860, 910, 990, 1 290.

Marquemos estes pontos sobre uma reta:



O primeiro que surge a nossa vista é que $x_{(10)} = 1\,290$ está muito afastado do resto da amostra. Na construção de um critério formal, nos ocorre ter um número que diga alguma coisa sobre a separação entre $x_{(n)}$ e as observações restantes, com respeito à dispersão da amostra. Por outra parte, é natural que o número buscado não dependa das unidades usadas para a medição, isto é, que se esse número nos diz que 1 290 é um *outlier* na amostra acima, a observação 5.29 também o seja na amostra:

4.79, 4.75, 4.91, 4.65, 4.99, 4.63, 5.29, 4.82, 4.86, 4.71, que é a resultante de aplicar uma transformação linear à anterior. Daí ser natural pensar em um número tal como N/D, onde N mede a separação de $x_{(n)}$ e D a dispersão amostral. Poderíamos usar para N algum dos seguintes: $N = x_{(10)} - x_{(9)} = 300$ (separação entre $x_{(10)}$ e seu vizinho mais próximo), $N = x_{(10)} - \bar{x}'$ (separação entre $x_{(10)}$ e a média da amostra sem esse ponto, que denotamos por $\bar{x}'$), isto é $= 500$; para D se poderia usar a amplitude da amostra sem $x_{(10)}$, isto é: $D = x_{(9)} - x_{(1)} = 360$, também é razoável usar $D = x_{(9)} - x_{(8)} = 80$ (separação entre os dois vizinhos mais próximos a $x_{(10)}$), ou também o desvio padrão da amostra sem $x_{(10)}, s' = 119$. Estas considerações nos sugerem usar como possíveis estatísticas para teste, as seguintes:

$$y(9,10;1,9) = (x_{(10)} - x_{(9)})/(x_{(9)} - x_{(1)})$$
$$y(9,10;8,9) = (x_{(10)} - x_{(9)})/(x_{(9)} - x_{(8)})$$
$$T' = (x_{(10)} - \bar{x}')/s'$$

que em nosso caso dão: 0.83, 3.75 e 4.20 respectivamente.

Dixon (1950, 1951), Likes (1966) e outros têm estudado o comportamento das estatísticas da forma:

$$y(r,s;\ p,q) = (x_{(s)} - x_{(r)})/(x_{(q)} - x_{(p)})$$

chamadas estatísticas de Dixon.

No caso que estamos estudando, também é razoável julgar o *outlier* pela razão entre $x_{(10)}$ e $x_{(9)}$ com respeito à amplitude da amostra completa, $(x_{(10)} - x_{(1)})$. Assim:

$$y(9,10;1,10) = (x_{(9)} - x_{(1)})/(x_{(10)} - x_{(1)}) = 300/660 = 0.45.$$

Mas para efeitos práticos esta estatística é a mesma que $y(9,10;1,9)$, pois:

$$\left(x_{(n)} - x_{(1)}\right) / \left(x_{(n)} - x_{(n-1)}\right) - \left(x_{(n-1)} - x_{(1)}\right) / \left(x_{(n)} - x_{(n-1)}\right) =$$
$$= \left(x_{(n)} - x_{(n-1)}\right) / \left(x_{(n)} - x_{(n-1)}\right) = 1.$$

Analogamente, a estatística T' é equivalente a:

$$T = (x_{(n)} - \bar{x})/s,$$

onde $\bar{x}$ e s são a média e o desvio padrão de toda a amostra. Com efeito:

$$(n-1)^2/(nT) - (n(n-2)) / \left((n-1)T'^2\right) = 1$$

As propriedades dos testes baseados sobre T' (ou T) foram estudadas por Pearson e Chandra Sekar (1936), Grubbs (1950), e outros autores.

Finalmente, outra estatística razoável seria:

$$\left(x_{(n)} - \bar{x}\right) / \left(x_{(n)} - x_{(1)}\right).$$

A respeito desta estatística, parece que pouco tem sido estudado, ainda que isto não seja grave, pois se suspeita que nada novo diria do ponto de vista prático.

Vejamos agora como influi na construção de um teste o tipo de distribuição que imaginamos se ajustar aos dados.

Suponhamos que os dados anteriores são observações de uma variável que, por experiências anteriores, se sabe que é uma gaussiana. O gráfico "Normal-Plot" de toda a amostra é mostrado na Figura 12.

Olhando esta figura, parece ser razoável supor a normalidade dos dados e que o 1 290 é um *outlier*. Vejamos agora o mesmo "Normal-Plot" sem esse ponto (ver a Figura 13). Agora é plenamente razoável supor que os dados $x_1, \cdots, x_9$ são ajustados por uma distribuição normal.

Consideremos um novo exemplo. A tabela seguinte mostra o número de dias de permanência, em uma sala de isolamento, em observação, de 92 pacientes em um hospital, antes de serem transferidos a uma sala principal (dados de Hoenin e Crotty (1958) citados em Barnett e Lewis (1984)):

Nº de dias:	1	2	3	4	5	6	7	8	9	10	11	21
Nº de pacientes:	11	18	28	8	12	5	5	1	1	1	1	1

A variável de nosso interesse será agora: Y = número de dias de permanência. Temos assim 92 observações. Vejamos um gráfico dos valores observados de Y (ver Figura 14). Olhando esta figura é evidente que $x_{(92)} = 21$ deveria ser considerado como um *outlier*. Pois bem, que critério poderia haver sido usado para identificá-lo? A suposição de que a distribuição da variável Y é gaussiana não é razoável, não obstante exista certa evidência para se supor que é do tipo Gamma com origem em zero. Com uma distribuição como essa o teste a usar deverá ser invariante por mudança de escala. Como no primeiro exemplo, pensamos em um teste baseado sobre uma estatística da forma N/D.

Recordemos que a distribuição Gamma(r,λ) está definida pela densidade:

$$f(x) = \frac{1}{\Gamma(r)}\lambda(\lambda x)^{r-1}e^{-\lambda x}$$

sendo sua esperança = r/λ e sua variância = r/λ^2. Se r é conhecido mas λ não é, então a dispersão da distribuição pode ser medida pela média amostral ou também pela soma das observações. Isto sugere que uma estatística útil para o que queremos poderia ser:

$$\frac{X_{(n)}}{\sum_{i=1}^{n} X_i}$$

que é equivalente, segundo vimos antes, a

$$\frac{X_{(n)}}{\sum_{i=1}^{n-1} X_i}.$$

Esta estatística seria inapropriada para casos onde se suspeita fortemente que a distribuição das observações é gaussiana. Por outro lado, as estatísticas de Dixon tal como

$$y(91,92;\ 1,91) = (x_{(92)} - x_{(91)})/(x_{(91)} - x_{(1)}),$$

são perfeitamente aplicáveis no caso de amostras de uma Gamma.

As considerações de índole geral, enunciadas nos exemplos anteriores, produzem uma ampla gama de estatísticas de testes de discordância que poderiam usar-se. Devemos agora perguntar quais são as adequadas para cada caso, como devem ser construídas, como poderia provar-se que satisfazem certos requisitos referentes a nível de significância e/ou potência e, finalmente, como julgar a performance dos testes que nos parecem possíveis de ser usados?

Em qualquer situação, se pode construir um teste sobre uma base meramente intuitiva. Isto é certo também para testes de *outliers* e, segundo temos visto, a intuição subjetiva tem desempenhado sempre um papel preponderante na eleição de um teste para detectar *outliers*. Porém, devemos ter presente, uma vez mais, as razões que vimos para esforçar-nos em conseguir uma base formal sólida sobre a qual construir

uma "teoria de testes de identificação de *outliers*". Esta base será mais necessária quando queremos comparar a performance de vários testes que nos parecem razoáveis de serem usados em uma determinada circunstância. Contudo, não veremos neste trabalho este ponto,e o leitor interessado pode encontrar muita coisa sobre o mesmo no livro de Barnett e Lewis (1984), onde também encontrará uma extensa lista de referências.

Terminaremos esta seção dando uma lista dos testes mais usados para o modelo univariado. Veremos, brevemente, como se poderia proceder para estabelecer seus níveis de significância e, finalmente, diremos algo sobre critérios para avaliar a performance de diferentes testes. Os testes que veremos aqui têm, todos eles, um significado intuitivo atraente e freqüentemente têm sido provadas, para alguns deles, diversas propriedades formais quando se aplicam a modelos apropriados.

Barnett e Lewis (1984) distinguem seis tipos básicos de estatísticas a serem usadas em testes de discordância. Dependendo do modelo que se considere, haverá algumas mais apropriadas que outras, por isso é bom tê-las todas presentes.

1) Estatísticas do tipo excesso/dispersão

São aquelas que estão definidas através do quociente entre a distância do possível *outlier*, a seu vizinho ou vizinhos mais próximos e a amplitude da amostra sem esse ponto, ou alguma medida de dispersão. Exemplos de estatísticas desta classe são:

$$(x_{(n)} - x_{(n-1)})/(x_{(n)} - x_{(2)})$$

que foi proposta por Dixon para examinar a possibilidade de que $x_{(n)}$ seja um *outlier*, sem considerar $x_{(1)}$. Outra proposta, especialmente útil para examinar *outliers* extremos em amostras gaussianas, é a sugerida por Irwin (1925):

$$(x_{(n)} - x_{(n-1)})/\sigma$$

Aqui se supõe que σ é conhecido, mas se poderia substituir σ por um estimador de escala no caso de não ser possível considerar σ conhecido.

2) Estatísticas do tipo amplitude/dispersão

Consistem em substituir nas do tipo anterior o numerador pela a amplitude (ver, por exemplo: David, Hartley e Pearson (1954), o Pearson e Stephens (1964)):

$$(x_{(n)} - x_{(1)})/\sigma$$

Como se usa a amplitude, não está claro qual deveria ser considerado *outlier*: se $x_{(n)}, x_{(1)}$ ou ambos. Isto é uma desvantagem.

3) Estatísticas do tipo desvio/dispersão

Estão formadas por modificações das anteriores para se resguardar da desvantagem recém-assinalada. Em lugar de usar a amplitude no numerador, usa-se alguma medida de distância entre o possível *outlier* e uma certa tendência central dos dados. Por exemplo, Grubbs (1950) propõe:

$$(\bar{x} - x_{(1)})/s.$$

Outra variante, sugerida por Halperin e outros (1955), substitui a amplitude pelo desvio máximo à média amostral:

$$(\max\{|x_{(i)} - \bar{x}|\})/s.$$

4) Estatísticas definidas por somas de quadrados

São expressas por quocientes entre somas de quadrados das diferenças entre as observações e médias das mesmas sobre subconjuntos da amostra e entre todas as observações e sua média. Por exemplo:

$$\frac{\sum_{i=1}^{n-2}\left(X_{(i)} - \bar{x}_{n,n-1}\right)^2}{\sum_{i=1}^{n}(X_i - \bar{x})^2}$$

onde $\bar{x}_{n,n-1} = \sum_{i=1}^{n-2} x_{(i)}/(n-2)$. Foi proposto por Grubbs (1950) para testar dois *outliers* nos extremos superiores da amostra.

5) Estatísticas baseadas sobre os momentos de grau alto

São aquelas do tipo "coeficiente de assimetria" e "kurtosis". Embora não tenham sido definidas para testar *outliers*, podem ser usadas para construir testes de discordância. Por exemplo, Ferguson (1961) define duas dessas estatísticas por:

$$\frac{\sqrt{n}\sum_{i=1}^{n}(X_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(X_i - \bar{x})^2\right)^{3/2}} \quad \text{e}$$

$$\frac{n\sum_{i=1}^{n}(X_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(X_i - \bar{x})^2\right)^2}$$

Outras estatísticas definidas fora do contexto de testes de *outliers* também são usadas para construir testes de discordância com boas propriedades. Por exemplo: a estatística W de Shapiro e Wilk (1965, 1972), também estudada em Shapiro, Wilk, e Chen (1968) e em Stephens (1978).

6) Estatísticas do tipo extremo/posição

São particularmente relevantes para examinar *outliers* quando se trabalha com distribuições hipotéticas da família Gamma. Um exemplo importante de tais estatísticas é:

$$x_{(n)}/\bar{x}$$

Têm sido analisadas em diversos trabalhos, por exemplo: Epstein (1960 a, b), Likes (1966), Lewis e Fieller (1979), Kimber e Stevens (1981) e Kimber (1982).

Em um trabalho de resenha, Grubbs (1969) dá exemplos ilustrativos sobre o uso de vários testes de discordância, aplicados a diversos conjuntos de dados reais.

Não podemos deixar de mencionar, ainda que brevemente, um assunto de importância quando se trabalha com testes de *outliers*. Trata-se do fenômeno chamado de "mascaramento". Consiste na tendência que afeta de modos diversos a todos os testes de discordância, de não detectar, em certas ocasiões, observações extremas como *outliers* quando em realidade o são. O termo "mascaramento" foi sugerido por Murphy (1951), ainda que pareça já ter sido discutido em Pearson e Chandra Sekar (1936); comentários adicionais sobre o tema podem ser encontrados em McMillan (1971) e em Tietjen e Moore (1972) que dão exemplos empíricos deste fenômeno.

Têm sido sugeridos também testes para verificar a presença de vários *outliers* simultaneamente. Um exemplo deles vimos no ponto 4. Uma forma diferente de abordar o problema é usar testes delineados para testar um *outlier* extremo, de maneira "seqüencial", isto é: deixando de lado o extremo detectado uma vez e repetindo o processo com as observações restantes. Um estudo detalhado de um procedimento desse tipo pode-se ver em McMillan (1971) e em McMillan e David (1971).

Na lista recém-dada aparecem estatísticas como a média amostral e diversas medidas de dispersão amostral. Uma dúvida razoável que nos surge é a seguinte: quando postulamos uma estatística que envolva essas medidas, para construir um teste, devemos ou não incluir os (ou o) suspeitos *outliers* no cálculo das médias e/ou dispersões? Pareceria que contamos com duas listas diferentes de estatísticas: as que incluem e as que excluem esses pontos suspeitos. Felizmente, na maioria dos casos práticos, os testes que resultam de usar uma estatística do tipo "inclusiva" e sua análoga "exclusiva" são equivalentes.

Com relação às bases formais que sustentam os testes assinalados mais acima, podemos dizer que, em geral, elas são: o princípio de máxima verossimilhança do quociente e o princípio de otimalidade local com respeito ao vício ou invariância. Uma vez mais, essa base formal dependerá da hipótese alternativa que se levante frente à hipótese nula de que não há *outliers* na amostra. O primeiro desses princípios parece ser o mais usado no caso de modelos mais complexos como o de regressão e no caso multivariado. Simplesmente, se considera como *outlier* aquele ponto cuja inclusão no cálculo do máximo da função de verossimilhança provoca a maior variação.

Certamente existem outros métodos de construção de testes de discordância. Por

exemplo, Kitagawa (1979) propõe o uso do "critério de informação de Akaike" (critério AIC), definido para uma amostra e um modelo paramétrico por:

AIC $=-2\log$ (máxima verossimilhança dos parâmetros do modelo suposto, sob a amostra) + 2 (número de parâmetros ajustados).

Esta proposta é bem atraente. Kitagawa em seu trabalho encarrega-se de destacar suas vantagens, entre elas: não é necessário fixar a priori um determinado nível de significância, o número de *outliers* resulta mesmo da aplicação do procedimento. Contudo, parece que este procedimento, em certos casos, apresentaria um defeito, digamos dual ao de "mascaramento", isto é, tenderia a classificar como *outliers* observações que não o seriam. Com efeito, Barnett e Lewis (1984) aplicaram esta proposta de Kitagawa aos dados de Herndon de 1846 sobre o semidiâmetro vertical do planeta Vênus (ver Seção 4) e obtiveram como valores discordantes não somente os pontos =1.4 e 1.01, mas também o 0.63, que dificilmente seria declarado *outlier* por qualquer analista que se baseie em sua experiência. Por este motivo, Barnett e Lewis acham duvidosa a utilidade prática deste procedimento.

A Estatística Não-Paramétrica também tem dito alguma coisa sobre o tema de tratamento de *outliers*. Alguns testes não-paramétricos para identificação de *outliers* e testes de discordância no caso univariado podem ser vistos em Walsh (1965), onde se faz referências a procedimentos já publicados nos anos 50. Técnicas não-paramétricas para acomodamento de *outliers* podem ser vistas em Walsh e Kelleher (1974). Contudo, estes procedimentos parecem de difícil justificação teórica e também são complicados de calcular. Por outro lado, existe um argumento filosófico contra o uso de procedimentos não-paramétricos para testar *outliers*. Com efeito, uma característica das técnicas não-paramétricas é que são úteis quando se faz um mínimo de suposições acerca da distribuição que gera os dados. Agora, os *outliers* são, por definição, observações que se julgam atípicas com respeito à maioria, isto é, são julgadas (explícita ou implicitamente) como geradas por distribuições diferentes da que gera a massa de dados. Assim, quando se tem pouca ou nenhuma base para estabelecer qual é esta última distribuição, mal pode esperar-se que se tenha uma base para julgar a "atipicidade" de uma ou mais observações.

Vejamos agora algo referente a como calcular o nível de significância de um teste de discordância. Suponhamos que tenhamos decidido usar uma certa estatística T para testar se uma configuração de dados, digamos x, tem ou não *outliers*. Suponhamos que o valor de T sobre x seja t, devemos decidir se x tem ou não *outliers*? Para responder a esta pergunta devemos fazer referência ao nível de significância de T, isto é, a $Sp(t)$ = probabilidade de que sob a distribuição nula (x não tenha *outliers*) T tome valores mais indicativos de discordância que t.

Na maioria dos casos, $Sp(t)$ será $P(T > t \mid H_0)$ ou $P(T < t \mid H_0)$.

Podemos então calcular $Sp(t)$ ou, uma vez fixado certo α, procurar em uma tabela da distribuição de T sob H_0 o valor t_α tal que $Sp(t_\alpha) = \alpha$ e compará-lo com t.

Mas freqüentemente, o valor $Sp(t)$ não está dado através de uma fórmula explícita que possibilite seu cálculo com precisão, ou não existem tabelas da distribuição de T sob H_0. Uma forma muito em voga atualmente, graças à crescente disponibilidade de computadores cada vez mais rápidos, é estimar esses valores por meio de simulações. Contudo, existem outros métodos que podem dar resultados mais rápidos e precisos que uma simulação, em certos casos. Podemos agrupar esses métodos em duas classes: métodos recursivos e métodos que usam desigualdades do tipo "desigualdades de Bonferroni". O leitor interessado pode encontrar referências a essas duas classes em Barnett e Lewis (1984).

Finalmente, trataremos brevemente o tema de critérios para avaliar a performance de testes de discordância. Certamente que uma medida importante será a $Sp(t)$ tratada anteriormente. Mas essa medida é pouco útil na maioria dos casos quando queremos efetuar comparações entre diversos testes. Assim, os critérios para efetuar tais comparações dependem das hipóteses alternativas que levantemos contra a hipótese nula de que não há *outliers* na amostra. Para simplificar a notação denotemos com H' essas hipóteses alternativas. Três probabilidades são relevantes para estabelecer um critério sobre performance de um teste:

P_1 = potência do teste (probabilidade de que sob H' o *outlier* seja identificado como observação discordante);

P_3 = probabilidade (sob H') de que o *outlier* seja "contaminante" e de que seja identificado como tal pelo teste;

P_5 = probabilidade de que quando o *outlier* seja "contaminante", o teste o identifique como discordante (isto é uma probabilidade condicional). A notação que usamos é a de Barnett e Lewis (1984). Depois de uma fundamentação baseada sobre uma classe particular de alternativa ("slippage alternative") esses autores afirmam que um bom teste deve estar caracterizado por: alta P_5, alta P_1 e baixa $P_1 - P_3$.

10 - TESTE DE DISCREPÂNCIA EM ÀMOSTRAS UNIVARIADAS NORMAIS

Como vimos, historicamente o problema de *outliers* surgiu na Astronomia, onde se procura determinar certos valores repetindo observações e medições. É razoável supor que a distribuição dos valores observados de uma mesma variável, nesses casos, seja aproximadamente normal. Não é surpreendente, assim, que a grande quantidade de trabalhos sobre *outliers* que se vem produzindo desde o Século XVIII se refira ao caso em que se supõe como distribuição hipotética dos dados uma normal. Somente nos últimos anos começaram a surgir trabalhos onde essa distribuição cede lugar a outras como a exponencial ou Gamma, em geral.

Neste trabalho nos limitaremos a apresentar alguns testes de discrepância quando se supõe como hipótese nula:

H; as observações são realizações de variáveis aleatórias $X_1, \cdots, X_n$ independentes, identicamente distribuídas com distribuição $N(\mu, \sigma)$ com μ e σ desconhecidos.

Devemos destacar que também existem testes de discrepância para os casos onde se supõe conhecido algum (ou os dois) desses parâmetros.

Do ponto de vista prático, é bom saber que as situações onde é razoável supor σ conhecido aparecem em problemas de controle de qualidade, nos quais experiências anteriores fornecem uma base razoável para o conhecimento preciso da variância do processo.

Situações onde são aplicáveis testes de discrepância com μ conhecido podem apresentar-se, por exemplo, quando se trata de analisar se existem *outliers* nas diferenças entre dois conjuntos de observações de uma mesma variável resposta, de modo que $\mu = 0$ é uma hipótese razoável.

Já os casos com μ e σ conhecidos têm um interesse mais limitado. Não obstante, poderiam apresentar-se em experiências realizadas para verificar a validade de uma tabela de números pseudo-aleatórios normalmente distribuídos. Também em certas experiências onde se busca ter uma base para tomar decisões sobre classificações taxonômicas ou antropológicas. Ou talvez em casos onde se deve emitir um certo juízo sobre uma situação nova em base a outras muito bem estudadas. Por exemplo, no caso de Mr. Hadlum (ver 2.1 da Seção 2), extensos estudos sobre a duração do período de gestação na mulher permitem supor valores razoavelmente precisos de μ e σ para a distribuição dessa variável (ainda que seja também razoável duvidar sobre a *gaussianidade* da mesma).

Outro caso interessante onde são úteis os testes de discrepância para normais com σ conhecido é o seguinte. Suponhamos que a distribuição nula é uma Poisson com parâmetro μ (desconhecido), então $x'_1, \cdots, x'_n$ com

$$x'_i \quad = \quad (x_i + 1/4)^{1/2}$$

se podem supor como realizações de variáveis aleatórias independentes e identicamente distribuídas com distribuição $N((\sigma)^{1/2}, 1/4)$. Claro que isto é aplicável também a outras distribuições com outras transformações.

De qualquer modo, como dissemos no princípio, nos limitaremos a considerar certos testes de discrepância relevantes para o caso normal com μ e σ desconhecidos. Estes testes, junto com vários outros para este caso e o de μ e/ou σ conhecidos e para outras distribuições diferentes da normal, são analisados com detalhe em Barnett e Lewis (1984). A notação que usamos neste trabalho é também a desse livro.

Usaremos a seguinte notação habitual:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$S^2 = \sum_{i=1}^{n}(x_i - \bar{X})^2$$

$$\bar{X}_n = \frac{1}{n-1}\sum_{i=1}^{n} X_{(i)}$$

$$\bar{X}_{n,n-1} = \frac{1}{n-2}\sum_{i=1}^{n-2} X_{(i)}$$

$$S_n^2 = \sum_{i=1}^{n-1}\left(X_{(i)} - \bar{X}_n\right)^2$$

$$S_{n,n-1}^2 = \sum_{i=1}^{n-2}\left(X_{(i)} - \bar{X}_{n,n-1}\right)^2$$

Tendo em conta que a distribuição normal é simétrica, qualquer teste delineado para testar se um ou vários valores extremos superiores são *outliers* é aplicável com modificações óbvias para o caso de valores extremos inferiores.

$N1$ – Teste de discrepância para um só *outlier* $x_{(n)}$ em uma amostra normal com média e variância desconhecidas.

Estatística do teste:

$$T_{N1} = \frac{X_{(n)} - \bar{X}}{s}$$

Nome: Desvio extremo da média internamente estudentizada.

Estatística equivalente:

$$\frac{S_n^2}{S^2} = 1 - \frac{n}{(n-1)^2}(T_{N1})^2$$

Relação de recorrência para calcular a densidade e/ou função de distribuição da estatística sob normalidade:

$$f_n(t) = \frac{n}{n-1}\left(\frac{n}{\pi}\right)^{1/2}\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-2}{2})}\left(1 - \frac{nt^2}{(n-1)^2}\right)^{\frac{n-4}{2}} \times$$

$$\times F_{n-1}\left(\left(\frac{n^2(n-2)t^2}{(n-1)\left((n-1)^2 - nt^2\right)}\right)^{1/2}\right);\ \frac{1}{\sqrt{n}} \le t \le \frac{n-1}{\sqrt{n}}$$

com

$$F_2(t) = \begin{cases} 0 & \text{se} \quad t < \dfrac{1}{\sqrt{2}} \\ 1 & \text{se} \quad t > \dfrac{1}{\sqrt{2}} \end{cases}$$

Desigualdade para calcular o nível de significância:

$$SP(t) \leq nP\left(t_{n-2} > \left(\frac{n(n-2)t^2}{(n-1)^2 - nt^2}\right)^{1/2}\right)$$

com igualdade quando

$$t \geq \left(\frac{(n-1)(n-2)}{2n}\right)^{1/2}$$

Propriedades do teste que rejeita H se T_{N1} é grande: $N1$ é o teste do quociente de verossimilhanças para testar H contra:

$H' = (n-1)$ observações provêm de $N(\mu, \sigma)$ e uma provém de $N(\mu + a, \sigma)$ com $a > 0$.

Com respeito a esta alternativa, tem a propriedade otimal de maximizar P_3 (ver final da Seção 9) entre os testes invariantes sob mudança de posição e escala.

É sensível com respeito ao fenômeno de "mascaramento" quando há mais de uma observação contaminante na amostra.

Não é conveniente usá-lo em forma consecutiva quando se suspeita que há vários *outliers*. Neste caso é preferível usar um teste como o $N15$ que será definido mais abaixo. Existem vários trabalhos onde se analisa e destaca esta desvantagem de $N1$. Por exemplo: Hawkins (1978), Prescott (1978), Hampel (1985).

Referências adicionais: Pearson e Chandra Sekar (1936), Dixon (1950, 1962), Grubbs (1950, 1969), Kudo (1956), Ferguson (1961 a, 1961 b), Quesenberry e David (1961), David e Paulson (1965), Stefansky (1971), McMillan (1971), Moran e McMillan (1973), Hawkins (1978), Hampel (1985).

$N2$ – Teste de discrepância para um só *outlier* $x_{(n)}$ ou $x_{(1)}$ em uma amostra normal com média e variância desconhecidas.

Estatística do teste:

$$T_{N2} = \max\left(\frac{\bar{X}_{(n)} - \bar{X}}{s}, \frac{\bar{X} - X_{(1)}}{s}\right)$$

Estatística equivalente:

$$\min\left(\frac{S_n^2}{S^2}, \frac{S_1^2}{S^2}\right) = 1 - \frac{n}{(n-1)^2}(T_{N2})^2$$

Desigualdade para calcular o nível de significância:

$$SP(t) \leq 2P(T_{N1} > t),$$

com igualdade para $t \geq ((n-1)(n-2)/(2n))^{1/2}$.

Propriedades do teste que rejeita H se T_{N2} é grande: $N2$ é o teste do quociente de verossimilhanças para testar H contra

$H' = (n-1)$ observações provêm de $N(\mu, \sigma)$ e uma provém de $N(\mu + a, \sigma)$ com $a \neq 0$.

Com respeito a esta alternativa, tem a propriedade otimal de maximizar P_3 (ver final da Seção 9) entre os testes invariantes sob mudança de posição e escala.

É sensível com respeito ao fenômeno de "mascaramento" quando há mais de uma observação contaminante na amostra, na mesma direção.

Ver também os comentários feitos sobre as propriedades do teste $N1$.

Referências adicionais: Kudo (1956), Ferguson (1961 a, 1961 b), Quesenberry e David (1961), Tietjen e Moore (1972), Prescott (1979).

$N3$ – Teste de discrepância para k *outliers*: $x_{(n-k+1)}, \cdots, x_{(n)}$ em uma amostra normal com média e variância desconhecidas.

Estatística do teste:

$$T_{N3} = \frac{X_{(n-k+1)} + \cdots + X_{(n)} - k\bar{X}}{s}$$

Nome: Soma de desvio com respeito à média internamente estudentizada.

Desigualdade para calcular o nível de significância:

$$SP(t) \leq \binom{n}{k} P\left(t_{n-2} > \left[\frac{n(n-2)t^2}{k(n-k)(n-1) - nt^2}\right]^{1/2}\right)$$

com igualdade para $t \geq \left(k^2(n-1)(n-k-1)/(nk+n)\right)^{1/2}$.

Propriedades do teste que rejeita H se T_{N3} é grande: $N3$ é o teste do quociente de verossimilhanças para testar H contra:

$H' = (n-k)$ observações provêm de $N(\mu, \sigma)$ e k provêm de $N(\mu + a, \sigma)$ com $a > 0$.

Com respeito a esta alternativa, tem a propriedade otimal de maximizar P_3 (ver final da Seção 9) entre os testes invariantes sob mudança de posição e escala.

Ver também os comentários feitos sobre as propriedades do teste $N1$.

Referências adicionais: Murphy (1951), Kudo (1956), Ferguson (1961 b), McMillan (1971).

*N*4 – **Teste de discrepância para** $k > 2$ *outliers*: $x_{(n-k+1)}, \cdots, x_{(n)}$ **em uma amostra normal com média e variância desconhecidas.**

Estatística do teste:

$$T_{N4} = \frac{S^2_{n-k+1,\cdots,n}}{S^2}$$

Nome: Soma de quadrados reduzida, dividida pela soma de quadrados total.

Propriedades do teste que rejeita H se T_{N4} é grande: $N4$ é o teste do quociente de verossimilhanças para testar H contra:

$H' = (n-k)$ observações provêm de $N(\mu, \sigma)$ e $k > 2$ provêm de $N(\mu + a_i, \sigma)$ com $a_i > 0$ e as $a_i's$ não necessariamente todas iguais, para $i = 1, \cdots, k$

Habitualmente, o número de *outliers* (k) tem que ser testado antes de ser eleito seja a "olho" ou bem como um parâmetro em um certo procedimento de processamento de dados.

Referências adicionais: Grubbs (1950, 1969), Dixon (1950), McMillan (1971), Tietjen e Moore (1972).

*N*6 – **Teste de discrepância para um par de** *outliers*, **um inferior e outro superior:** $x_{(1)}, x_{(n)}$ **em uma amostra normal com média e variância desconhecidas.**

Estatística do teste:

$$T_{N6} = \frac{X_{(n)} - X_{(1)}}{s}$$

Nome: amplitude estudentizada internamente.

Desigualdade para calcular o nível de significância:

$$SP(t) \leq n(n-1)P\left(t_{n-2} > \left[\frac{(n-2)t^2}{2n-2-t^2}\right]^{1/2}\right)$$

com igualdade quando

$$t \geq \left[\frac{3}{2}(n-1)\right]^{1/2}$$

Propriedades do teste que rejeita H se T_{N6} é grande: $N6$ tem boas propriedades referentes à potência quando a alternativa é uma distribuição simétrica mas sua performance é pobre quando a alternativa é uma assimétrica.

Referências adicionais: David, Hartley e Pearson (1954), Pearson e Stephens (1964), Shapiro, Wilk e Chen (1968).

*N*7 – **Teste de discrepância para um só *outlier* $x_{(n)}$ em uma amostra normal com variância desconhecida (a média poderia ser conhecida ou não).**

Estatística do teste:

$$T_{N7} = (x_{(n)} - x_{(n-1)})/(x_{(n)} - x_{(1)})$$

Nome: Estatística r_{10} de Dixon.

Propriedades do teste que rejeita H se T_{N7} é grande: é eficiente principalmente quando há no máximo uma observação contaminante e esta é $x_{(n)}$. É vulnerável ao efeito de "mascaramento" quando $x_{(n-1)}$ e/ou $x_{(1)}$ é *outlier*.

Referências adicionais: Dixon (1950, 1951), Ferguson (1961 a, 1961 b).

*N*8 – **Teste de discrepância de "duas caudas" para um *outlier* extremal em uma amostra normal com variância desconhecida (a média poderia ou não ser conhecida).**

Estatística do teste:

$$T_{N8} = \max\left(\frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(1)}}, \frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}}\right)$$

Desigualdade para calcular o nível de siginificância:

$$SP(t) \leq 2P(T_{N7} > t)$$

com igualdade quando $t \geq 1/2$. Propriedades do teste que rejeita H se T_{N8} é grande: são similares as já destacadas para $N7$ pois não é mais que o teste de "duas caudas" correspondente a $N7$.

Referência principal: King (1953).

*N*13 – **Teste de discrepância de "duas caudas" para um par de *outliers* em um dos extremos em uma amostra normal com variância desconhecida (a média poderia ou não ser conhecida).**

Estatística do teste:

$$T_{N13} = \max\left(\frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(3)}}, \frac{X_{(3)} - X_{(1)}}{X_{(n-2)} - X_{(1)}}\right)$$

Nome: Estatística r_{22} de Dixon.

Propriedades do teste $N13$ que rejeita H se T_{N13} é grande: protege contra o efeito de "mascaramento" de *outliers* situados nos dois pontos extremos superiores ou extremos inferiores da amostra.

Referências adicionais: Dixon (1950, 1951), Ferguson (1961 b).

$N14$ – **Teste de discrepância para um ou mais *outliers* extremais em uma amostra normal com média e variância desconhecidas.**

Estatística do teste:

$$T_{N14} = \sqrt{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^{3/2}}$$

Nome: assimetria amostral.

O teste $N14$ baseado nesta estatística rejeita o extremo superior ou o extremo inferior conforme o sinal, + ou - respectivamente, de T_{N14}. Para mais de um *outlier* se aplica consecutivamente. Propriedades do teste $N14$: é o localmente ótimo entre os testes invariantes, dado o tamanho n, para testar H contra:

$H' = (n-k)$ observações provêm de $N(\mu, \sigma)$ e $[n/2] > k > 1$ provêm de $N(\mu + a_i, \sigma)$ com $a_i > 0$ e as a_i's não necessariamente todas iguais, para $i = 1, \cdots, k$.

Sua potência é quase tão boa quanto a de $N1$ quando $k = 1$. Também tem boa potência quando a alternativa é

$H' =$ as n observações têm distribuição G, e se supõe que G é Cauchy ou bem que é uma log-normal. Também são válidas as advertências feitas para $N1$ com respeito ao "mascaramento".

Referências adicionais: Ferguson (1961 a, 1961 b), Shapiro, Wilk e Chen (1968).

$N15$ – **Teste de discrepância para um ou mais *outliers* extremais qualquer que seja sua direção em uma amostra normal com média e variância desconhecidas.**

Estatística do teste:

$$T_{N15} = n\frac{\sum_{i-1}^{n}(X_i - \bar{X})^4}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

Nome: kurtosis amostral.

O teste $N15$ baseado nesta estatística rejeita o extremo superior ou o extremo inferior que esteja mais afastado da média amostral. Para mais de um *outlier* se aplica

consecutivamente. Propriedades do teste $N15$: é o localmente ótimo entre os testes invariantes e não viciados de tamanho n, para testar H contra:

$H' = (n-k)$ observações provêm de $N(\mu,\sigma)$ e $k > 1, k/n < 0.21$, provêm de $N(\mu+a_i,\sigma)$ com $a_i \neq 0$ e as a_i's não necessariamente todas iguais, para $i = 1,\cdots,k$.

$N15$ também é o localmente ótimo entre os testes invariantes dado o tamanho n, para testar H contra

$H' = (n-k)$ observações provêm de $N(\mu,\sigma)$ e $k > 1$ provêm de $N(\mu, b_i\sigma)$ com $b_i > 1$, para $i = 1,\cdots,k$.

Sua potência é quase tão boa quanto a de $N2$ quando $k = 1$ e a alternativa é como a vista em $N2$. Mas sua potência é notavelmente maior quando $k > 2$, principalmente para amostras de tamanho $n < 20$. Tem a vantagem de ser robusto contra possíveis "mascaramentos". Ele é também conveniente para ser usado em forma consecutiva quando se suspeita de que existem mais de um *outlier*.

Referências adicionais: Ferguson (1961 a, 1961 b), Shapiro, Wilk e Chen (1968), Hampel (1985).

$N17$ – Teste de discrepância de "duas caudas" para testar a presença de um número indefinido de contaminantes em uma amostra normal com média e variância desconhecidas.

Estatística do teste:

$$T_{N17} = \frac{\left(\sum_{i=1}^{[n/2]} a_{n,n-i+1}(X_{(n-i+1)} - X_{(i)})\right)^2}{S^2}$$

onde $[n/2]$ denota a parte inteira de $n/2$, e os $a_{n,j}$'s são certas constantes definidas e tabuladas em Shapiro e Wilk (1965).

Nome: W-estatística de Shapiro e Wilk.

O teste $N17$ rejeita a hipótese H quando T_{N17} é pequeno em favor da alternativa geral:

H' = existem observações contaminantes na amostra.

Referências adicionais: Shapiro e Wilk (1965), Shapiro, Wilk e Chen (1968), Chen (1971).

$N18$ – Teste de discrepância para um só *outlier*: $x_{(n)}$ ou $x_{(1)}$ em uma amostra normal com média e variância desconhecidas usando um estimador robusto para o desvio padrão.

Estatística do teste:

$$T_{N18} = \max\left(\frac{X_{(n)} - \bar{X}}{s_b}, \frac{\bar{X} - X_{(1)}}{s_b}\right)$$

onde

$$s_b^2 = \frac{V}{W}\sqrt{n}$$

com

$$V = \left(\sum_{|u_i|<1} (X_i - m)(1 - u_i^2)^4\right)^{1/2}$$

$$W = \max\left(1, -1 + |\sum_{|u_i|<1} (1 - u_i^2)(1 - 5u_i^2)|\right)$$

$$u_i = (X_i - m)/(9.\quad \text{Mediana}\quad (|x_1 - m|, \cdots, |x_n - m|))$$

e sendo m a mediana amostral.

Propriedades do teste $N18$ que rejeita H se T_{N18} é grande: Se trata de um procedimento pragmático. A estatística s_b^2 está motivada em um estimador de escala proposto em Mosteller e Tukey (1977) baseado em um estimador da variância assintótica do M-estimador definido pelo tipo biquadrada (ver Seção 8). As observações que estão afastadas de m em mais de 6 são ignoradas. Se as $u_i's$ são pequenas, s_b^2 é aproximadamente igual a $\sum_{i=1}^n (x_i - m)^2/(n-1)$. Iglewicz e Martinez (1982) fazem algumas comparações empíricas entre este teste e os testes $N2$, $N14$ e $N15$. Mesmo que a potência deste teste nos casos comparados não tenha revelado defeitos sérios, este procedimento é laborioso e é necessário realizar investigações mais detalhadas.

HS – Teste de discrepância de "duas caudas" para testar a presença de um número indefinido de contaminantes em uma amostra normal com média e variância desconhecidas usando o M-estimador de Huber.

Para $k > 0$ seja ψ_k a função definida por:

$$\psi_k(x) = xI_{[-k,k]}(x)$$

Seja t = um certo estimador robusto para μ, s = um certo estimador robusto para σ.

Definição do teste: considerar como *outliers* contaminantes todas as observações $x_i's$ que estão fora do intervalo $(t - ks, t + ks)$.

A constante k se determina sobre a base do nível de confiança, α, do teste sob a hipótese nula H: " Não existem *outliers* na amostra ". Isto é, dado α,

k se define por:

$$P_H\left(\sharp\{i: \frac{|x_i - t|}{s} < k\} < n\right) \leq \alpha, \text{ onde } \sharp \text{ denota a cardinalidade do conjunto}$$

Versão "one-step":

$$t = \text{Mediana } (X_1, \cdots, X_n),$$
$$s = \text{Mediana } (|X_1 - t|/0.6745, \cdots, |X_n - t|/0.6745).$$

Versão "solução iterativa": t é a solução de

$$\sum_{i=1}^{n} \psi_k\left(\frac{x_i - t}{s}\right) = 0$$

onde s é como acima.

Versão "Huber Proposal-2": (t, s) é a solução do sistema:

$$\sum_{i=1}^{n} \psi_k\left(\frac{x_i - t}{s}\right) = 0$$
$$\sum_{i=1}^{n} \psi_k^2\left(\frac{x_i - t}{s}\right) = n\beta$$

onde $\beta = \int \psi_k^2(x)\varphi(x)dx$, com $\varphi(x)$ a densidade da $N(0,1)$.

Como vemos, se trata de um teste bastante útil pois não precisa que postulemos antecipadamente o número de *outliers*.

O comportamento deste teste foi extensivamente estudado por Schweingruberem em sua tese de doutorado apresentada no ETH, Suíça e, lamentavelmente não foi publicada nenhuma tradução da mesma (está em alemão). Como alternativa, alguma coisa podemos ver no trabalho de Hampel (1985) e de Hampel e outros (1986).

11 - COMPARAÇÃO ENTRE ESTIMADORES ROBUSTOS E PROCEDIMENTOS DO TIPO "TESTE DE DISCREPÂNCIA + REJEIÇÃO DE *OUTLIERS* + MÉDIA AMOSTRAL"

Nesta Seção nos limitaremos a sintetizar resumidamente o trabalho de Hampel (1985). Sem dúvida trata-se de um trabalho pioneiro no tema de relação entre *outliers* e *robustez*. Trata-se, também de um dos primeiros trabalhos onde se enfatiza a utilidade do conceito de "ponto de ruptura" tanto na teoria como na prática. Com efeito, esse

conceito, convenientemente definido, aplicado aos testes de *outliers*, fornece uma base formal para explicar o fenômeno de "mascaramento" de *outliers*.

Essencialmente, Hampel (1985) analisa as variâncias "Monte Carlo" de vários procedimentos estatísticos que consistem em tomar a média das observações que permanecem na amostra depois de haver retirado os *outliers* detectados conforme um certo teste de discrepância. Para abreviar, designamos estes procedimentos "rejeição + estimação". Ele compara as variâncias destes com as correspondentes a certos estimadores robustos; define, sem entrar em formalismos, o conceito de "ponto de ruptura" e calcula esta magnitude para essas duas classes de técnicas de tratamento de *outliers* em estimação.

Este trabalho de Hampel, por sua vez, está baseado no "estudo de Princeton" (Andrews e outros (1972)) que foi continuado em Gross e Tukey (1973). Hampel testou 32 variantes de seis tipos de procedimentos do tipo "rejeição + estimação". As variantes foram criadas fixando diversos valores críticos nos correspondentes testes de discrepância, e por versões *one step* e iterativas. As do tipo *one step* rejeitam no máximo um *outlier*, as iterativas testam cada vez a observação mais distante e recalculam o teste até que não haja mais rejeições.

Neste trabalho os testes que se usaram foram os recomendados para alternativas que permitiam *outliers* em ambos os lados da amostra. Os seis procedimentos "rejeição + estimação" usados foram baseados sobre os seguintes testes de discrepância (a notação é a da Seção anterior): $N15$, $N2$, $N6$, $N13$, $N17$ e HS. Foram simuladas amostras de tamanho 20 das seguintes distribuções: $N = N(0,1)$, $5\%3N$, $5\%10N$, $10\%10N$, t_3, $25\%10N$ e Cauchy; onde, para cada x e y, se denota por $x\%yN$ a distribuição tal que $20x/100$ das observações provêm da $N(0,y)$ e as restantes da $N(0,1)$. O número de replicações efetuadas foi entre 640 e 1000, analogamente ao que se fez em Andrews e outros (1972).

Na tabela, que veremos mais adiante, estão somente alguns dos resultados obtidos. Nela se mostram os valores de 20 vezes VMC, onde VMC é a variância Monte Carlo, para cada um dos estimadores. Na última coluna se dão os valores dos "pontos de ruptura", cuja definição veremos em seguida. Também se mostram os valores para 3 estimadores robustos: mediana amostral (é o denotado por 50% na Tabela), $H15$ e $25A$; sendo:

$H15$, o M-estimador definido pela ψ de Huber com $k = 1.5$ e estimador da escala definido simultaneamente como na "Huber Proposal 2" de Huber (1964);

$25A$, o M-estimador definido em Andrews e outros (1972) como o M-estimador dado pela ψ de Hampel (ver Seção 8) com:

$$a = 2.5s \approx 1.7\hat{\sigma},$$
$$b = 4.5s \approx 3.0\hat{\sigma},$$
$$c = 9.5s \approx 6.4\hat{\sigma},$$

com s como o estimador robusto de escala que temos chamado de MAD e $\hat{\sigma}$ o desvio padrão amostral.

A seguir nos ocupamos de formalizar um pouco o conceito de "ponto de ruptura" tanto para os estimadores do tipo "rejeição + estimação" como para os robustos.

Naturalmente, para os procedimentos "rejeição + estimação", que estão baseados sobre algum teste de discrepância, dizer que dão informação útil ou que não tenham alcançado seu "ponto de ruptura", quando a amostra tem uma proporção δ de contaminantes, significa que todos os *outliers* (do tipo contaminante) suficientemente afastados tenham sido detectados pelo correspondente teste de discrepância. Hampel mostra que, para um teste de discrepância dado, sempre existem configurações de amostras com *outliers* arbitrariamente afastados, tais que esse teste não detecta, e assim esses pontos são incluídos no cálculo da média como observações "boas".

As definições de "ponto de ruptura" que veremos a seguir estão baseadas essencialmente na de Huber (1984), já Hampel e outros (1986) expõem uma ligeiramente diferente. Ainda que não haja sido provado efetivamente, é de supor-se que, nos casos deste estudo Monte Carlo, ambas versões dão resultados similares, isto é, que não mudariam qualitativamente a avaliação da performace das técnicas em questão.

Primeiramente, definimos o que se chamou anteriormente "amostra com outliers substitutivos". Seja $\mathbf{y} = (y_1, \ldots, y_n)'$ um vetor n-dimensional (uma amostra de tamanho n), $0 \leq m < n$ um inteiro, $\mathbf{z} = (z_1, \ldots, z_m)'$ um vetor m-dimensional, J uma função injetiva definida sobre $\{1, \ldots, m\}$ com valores em $\{1, \ldots, n\}$. Seja $\mathbf{x} = (x_1, \ldots, x_n)$ definido por:

$$x_j = y_j \text{ se } j \text{ não pertence a } \{J(1), \ldots, J(m)\},$$
$$x_j = z_k \text{ se } j = J(k) \text{ para } 1 \leq k \leq m.$$

A $\mathbf{x}$ é chamada (m/n)-contaminação (por substituição) de $\mathbf{y}$.

A distinção "por substituição" se faz necessária quando se quer trabalhar com outros tipos de contaminações, por exemplo, quando a amostra observada, $\mathbf{x}$, é a $\mathbf{y}$ mais outros pontos "espúrios" . Por este motivo Huber (1984) chama "corrupção" de $\mathbf{y}$ à $\mathbf{x}$, abarcando com esta denominação qualquer tipo de contaminação. Neste trabalho consideramos somente as contaminações por substituição que parecem ser as mais freqüentes.

Para cada $m \leq n$ seja

$$C(m/n, \mathbf{y}) = \{\mathbf{x} \mid \mathbf{x} \text{ é uma (m/n)-contaminação de } y\}.$$

Consideremos agora um teste de discrepância, digamos NT, baseado sobre uma estatística T que rejeita a hipótese nula "não há *outliers* na amostra" se T é maior que certa constante k (valor crítico). Seja agora

$$c(m/n, \mathbf{y}, NT) = \sup\{T(\mathbf{x}) : \mathbf{x} \in C(m/n, \mathbf{y})\}.$$

Definimos como "ponto de ruptura" do teste NT na amostra y a

$$\delta^*(\mathbf{y}, NT) = \sup\{m/n : c(m/n, \mathbf{y}, NT) > k, \quad m = 1, \ldots, n\}.$$

Fica assim definido o "ponto de ruptura" para um teste de discrepância. Vejamos agora a definição para um estimador.

Seja t um estimador de um certo parâmetro, pensemos na média da distribuição para fixar idéias. Se define "vício máximo de t em $\mathbf{y}$" a

$$b(m/n, \mathbf{y}, t) = \sup\{|t(\mathbf{y}) - t(\mathbf{x})| : \mathbf{x} \in C(m/n, \mathbf{y})\}.$$

Se chama "ponto de ruptura" de t em $\mathbf{y}$ para *outliers* contaminantes do tipo substitutivo a

$$\epsilon^*(\mathbf{y}, t) = \sup\{m/n : b(m/n, \mathbf{y}, t) < \infty, \quad m = 1, \ldots, n\}$$

é possível provar que no caso de nosso interesse, estimação da média de uma normal, se t é o estimador da média, derivado de um processo "rejeição + estimação" cujo teste de discrepância é NT, então os valores δ^* e ϵ^* são iguais.

Também é possível ver que, geralmente, os valores dos pontos de ruptura não dependem de $\mathbf{y}$ e sim de n e do estimador. É interessante analisar o comportamento assintótico (quando n cresce) do ponto de ruptura. Na realidade, as primeiras definições de ponto de ruptura encontradas na literatura estatística são do tipo "assintóticas". Se baseavam no funcional que define, geralmente, um estimador (recordar o que dizíamos sobre este tema na Seção 7).

Em seu trabalho, Hampel dá, com algum detalhe, as diretrizes gerais que permitem calcular os pontos de ruptura definidos acima. Para termos alguma idéia sobre os mesmos, vejamos como proceder no caso da kurtosis amostral (estimador T_{N15}). Neste caso, e em muitos dos outros estimadores, se pode ver que a configuração de *outliers* mais desfavorável para o estimador é aquela que tem os *outliers* muito próximos um do outro e todos eles muito afastados dos "dados bons". Para facilitar as contas, imaginemos que os "dados bons" estejam muito próximos a zero enquanto que os "maus" estão concentrados a uma distância grande, digamos x. Resumindo, sejam:

$$\begin{aligned} &\mathbf{y} = (0, \ldots, 0)', \quad m < n, \mathbf{x} \quad \text{dado por:} \\ &x_i = 0 \quad \text{para} \quad i = 1, \ldots, n - m; \\ &x_i = x \quad \text{para} \quad i = n - m + 1, \ldots, n \end{aligned}$$

Temos então:

$$\bar{x} = -(m/n)x,$$

$$(x_j - \bar{x}) = \begin{cases} -(m/n)x & \text{se} \quad j = 1, \ldots, n - m \\ (1 - (m/n))x & \text{se} \quad j = n - m + 1, \ldots, n \end{cases}$$

Para simplificar a notação ponhamos $d = (m/n)$, então resulta:

$$\sum_{j=1}^{n}(x_j - \bar{x})^4 = (n-m)d^4x^4 + m(1-d)^4x^4$$

$$ns^4 = (n^3/(n-1)^2)x^4d^2(1-d)^2$$

e daí:

$$T_{N15} = \frac{((n-1)/n)^2(1-3d+3d^2)}{d(1-d)} \approx \frac{1-3d+3d^2}{d(1-d)}$$

Logo, se k é o nível crítico do teste $N15$, a amostra contaminada será detectada se

$$\frac{1-3d+3d^2}{d(1-d)} = \frac{1}{d(1-d)} - 3 > k$$

em definitivo, resulta que o ponto de ruptura de $N15$ com o valor crítico k é o δ tal que

$$\frac{1}{\delta^*(1-\delta^*)} - 3 = k$$

Por exemplo: o valor crítico correspondente ao nível de significação $1-\alpha = 99\%(\alpha = 1\%)$ é aproximadamente $k = 5$, então $\delta^* = 14.7\%$. Isto quer dizer que uma proporção de *outliers* próximos uns dos outros e muito afastados dos dados "bons", menor que uns 15%, será detectada, mas já uma proporção maior não o será. Usando o conceito de "mascaramento", diríamos que uma proporção maior do 15% desse tipo de *outliers* fica "mascarada". Notemos que a noção precisa de "ponto de ruptura" formaliza o conceito vago de "mascaramento".

Vejamos agora alguns resultados do trabalho de Hampel.

TABELA 20 VEZES VMC E PONTOS DE RUPTURA

Robustos

estim	N	5%3N	5%10N	10%10N	t_3	25%10N	Cauchy	δ^*
50%	1.49	1.52	1.56	1.80	1.82	2.5	2.9	50%
$H15$	1.04	1.16	1.21	1.50	1.71	4.0	5.7	26%
25A	1.05	1.16	1.13	1.26	1.67	2.1	3.7	50%

Rejeição + estimação

estim	k	α	N	5%3N	5%10N	10%10N	t_3	25%10N	Cauchy	δ^*
$N15$	3.5	10%	1.07	1.17	1.16	1.26	1.93	2.7	4.7	19%
$N2$	2.56	10%	1.05	1.16	1.13	1.28	1.95	7.6	6.0	13%
$N6$	4.32	10%	1.04	1.19	1.16	4.76	2.21	22.6	28.9	5%
HS	3.5	10%	1.04	1.20	1.14	1.31	1.85	2.4	4.2	50%
$N13$	.45	10%	1.04	1.20	1.15	1.29	2.05	13.3	7.7	13%
$N17$	.92	10%	1.27	1.30	1.24	1.40	1.84	2.1	3.5	48%

Vamos finalizar esta Seção resumindo as principais conclusões que podem ser extraídas deste trabalho de Hampel:

1) $N6$ não é capaz de rejeitar um *outlier* distante.

2) $N2$ não é capaz de rejeitar dois *outliers* distantes em amostras de tamanho 20.

3) Os testes que produziram os melhores procedimentos do tipo "rejeição + estimação" foram $N17$ (teste de Shapiro-Wilk) e o "Huber-skip".

4) $N15$ (a kurtosis amostral) é moderadamente confiável.

5) Todos os resultados podem ser explicados em forma precisa usando-se o conceito de "ponto de ruptura". Simplesmente: as variâncias "explodem" quando a fração de *outliers* ultrapassa o "ponto de ruptura".

6) O "ponto de ruptura" formaliza a idéia de "mascaramento" e pode-se usá-lo em seu lugar. Uma amostra com uma proporção de *outliers* maior que a de "ponto de ruptura" fará com que o teste não seja capaz de detectar alguns de tais *outliers*.

7) Também o "ponto de ruptura" está relacionado estreitamente com a potência de um teste de discrepância sob a alternativa de *outliers* de tipo contaminante. Se a proporção de *outliers* for menor que o ponto de ruptura, a potência tenderá a 1 quando a massa de *outliers* tende a afastar-se "infinitamente" da massa de dados.

8) Dado que os procedimentos para manejo de *outliers* buscam fornecer um procedimento que assegure uma confiabilidade global, parece que o critério baseado sobre a maximização do ponto de ruptura é mais relevante que outros critérios baseados sobre outras propriedades de otimalidade, tais como a eficiência assintótica, por exemplo.

9) Uma característica negativa dos procedimentos "rejeição + *outlier*" (com notável

exceção do baseado sobre HS) é

a rápida perda de eficiência, mesmo sob configurações amostrais com uma proporção de *outliers* menor que o ponto de ruptura.

10) Agora uma observação favorável aos procedimentos "rejeição + estimação": se os *outliers* estão bem afastados e sua proporção é menor que o ponto de ruptura, então esses procedimentos são tão bons como os melhores estimadores robustos.

11) Se os *outliers* estão mais ou menos "perto" dos dados "bons", como sucede no caso da hipótese alternativa ser t_3, e sua proporção ser menor que o ponto de ruptura, então sua eficiência será muito menor que a dos M-estimadores com ψ de tipo Huber e mais ainda que a correspondente aos definidos por ψ do tipo "redescendente".

12) Possivelmente a principal debilidade dos procedimentos "rejeição + estimação" consista em sua pobre performance quando os *outliers* estão concentrados perto da fronteira da região de rejeição do teste. Com efeito, como esses procedimentos são drásticos – simplesmente aceitam ou rejeitam certas observações –, quando uma delas está perto dessa fronteira, uma ligeira modificação (produzida por arredondamento ou leve falha no instrumento de medição, por exemplo) poderia fazê-la rejeitada ou aceita com igual probabilidade. Isto poderia conduzir a um comportamento errático do estimador final, inclusive com variância que não tendesse a zero quando n tende ao infinito.

13) Moral final: em análise de dados não basta usar regras formais de rejeição de *outliers* a fim de identificá-los em uma amostra.

MICROTAB - versão 2.1

*** UNIVARIADA ***

GRÁFICO PROBABILÍSTICO NORMAL

Variável: Var1 - Variável 1

Casos válidos 10
Mínimo: 6.3000 E + 02
Máximo: 1.2900 E + 03

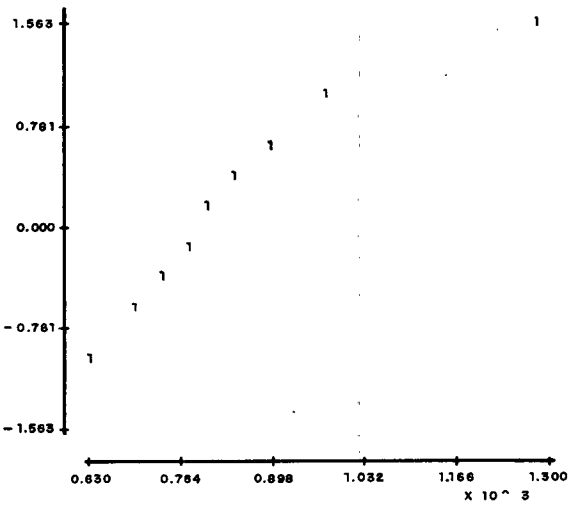


FIGURA 12

MICROTAB - versão 2.1

*** UNIVARIADA ***

GRÁFICO PROBABILÍSTICO NORMAL

Variável: Var1 - Variável 1

Casos válidos 9
Mínimo: 6.3000 E + 02
Máximo: 9.9000 E + 02

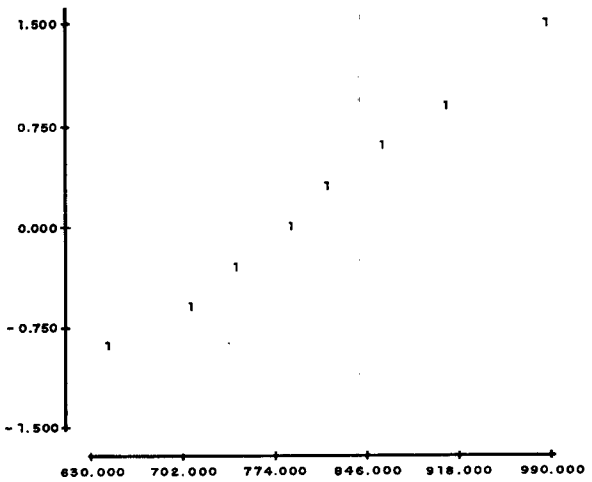


FIGURA 13

MICROTAB - versão 2.1

*** UNIVARIADA ***

GRÁFICO PROBABILÍSTICO NORMAL

Variável : var 1 - variável 1

Casos válidos : 92
Mínimo : 1.0000 E + 00
Máximo : 2.1000 E + 01

* equivale a 10 ou mais pontos

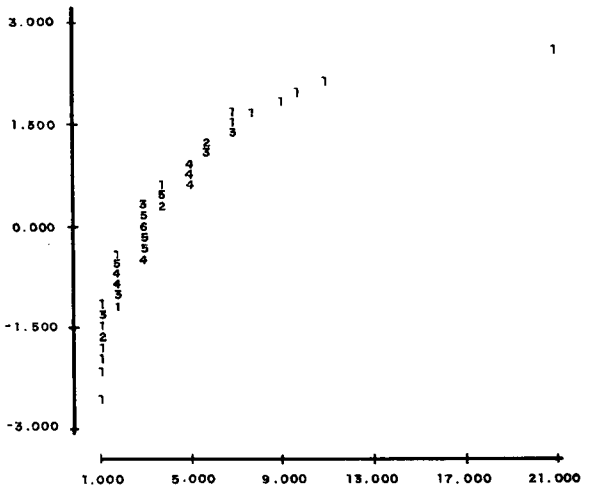


FIGURA 14

BIBLIOGRAFIA

ANDREWS, D. F. et al. *Robust estimates of location: surveys and advances.* Princeton, New Jersey : Princeton University Press, 1972.

BARNETT, V., LEWIS, T. Outliers in statistical data 2.ed. New York : Wiley, 1984.

CHEN, E. H. The power of the Shapiro–Wilk W–Test for normality in samples from contaminated distribution. *JASA*, v.66, n.336, p.760–762, 1971.

DAVID, H. A., HARTLEY, H.O. PEARSON, E.S. The distribution of the ratio in a single normal sample, of range to standard deviation. *Biometrika*, v.41, p.482–493, 1954.

DAVID, H.A., PAULSON, A.S. The performance of several tests for outliers. *Biometrika*, v.52, p.429–436, 1954.

DIXON, W.J. Analysis of extreme values. *Ann. of Math Stat.*, v.21, n.4, p. 488–506, 1950.

———— Ratios involving extrem e values. *Ann. of Math. Stat.*, v.22, n.1, p.68–78, 1951.

DIXON, W. J. Rejection of observations. In: CONTRIBUTIONS to order statistics. New York : Wiley, 1962.

EPSTEIN, B. Tests for the validity of the assumption that the underlying distribution of life is exponential. *Technometrics*, v.2, part. I, p.83–101, 1960.

————. Tests for the validity of the assumption that the underlying distribution is exponential. *Technometrics*, v.2, part. II, p. 167–183, 1960.

FERGUSON, T. S. On the rejection of outliers. In: BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 4, 1961. Berkeley. Proceedings ... Berkeley, University of California Press, 1961. v.1. p. 253–287.

———— Rules for rejection of outliers. *Rev. Inst. Int. Stadist.*, v. 29, p. 29–43.

GROSS, A. M., TUKEY, J. W. *The estimators of the Princeton robustness study.* Princeton, New Jersey : Princeton University, Departament of Statistics, 1973. (Technical report n. 38, ser. 2).

GRUBBS, F. E. Sample criteria for testing outlying observations. *Ann. of Math. Stat.*, v.21, n.1, p.27–58, 1950.

———— Procedures for decting outlying observations in samples. *Technometrics*, v.11, n.1, p. 1–21, 1969.

HAMPEL, F. R. The breakdown points of the mean combined with some rejection rules. *Technometrics*, v.27, n.2, p. 95–107, 1985.

———— et al. *Robust statistics: the approach based on influence functions.* New York : Wiley, 1955.

HALPERIN, M. et al. Tables of percentage points for the studentized maximum absolute deviate in normal samples. *J. Amer. Statist. Assoc.*, v. 50, n. 269, p. 185–195, 1955.

HAWKINS, D. M. Analysis of three testes for one or two outliers. *Statistica Weerlandica*, v.32, p.137–148, 1978.

HOENIG, J., CROTTY. *International J. Social Psychiatry*, v.3, n. 110, p. 260–277, 1958.

HUBER, P. J. Robust estimation of a location parameter *Ann. Math. Statist.*, v.35, n.1, p. 73–101, 1964.

———— Finite sample breakdown of M – and P–estimators. *Ann. Statist.*, v.12, n.1, p. 119–126, 1984.

IGLEWICZ, B., MARTINEZ, J. *Outlier detection using robust measures of scale.* Philadelphia : Temple University, Dept. of Statistics, School of Business Administration, 1982. (Techical report (theory); 12).

KIMBER, A. C. Tests for many outliers in an exponential sample. *Applied statistics*, v.31, n.3, p. 263–271, 1982.

KIMBER, A. C., STEVENS, H. J. The null distribution of a test for two upper outliers in an exponential sample. *Applied statistics*, v. 30, n. 2, p. 153–157, 1981.

KING, E. P. On some procedures for the rejection of suspected data. *JASA*, v.48, p. 531–533, 1953.

KITAGAWA, G. On the use of AIC for the detection of outliers. *Technometrics*, v.21, n. 2, p. 193–199, 1979.

KUDO, A. On the testing of outlying observations. *Sankyā*, v. 17, part.1, p. 67–76, 1956

LEWIS, T., FIELLER, N. R. J. A recursive algorithm for null distributions for outliers: I Gamma samples. *Technometrics*, v. 21, n.3, p. 371–376, 1979.

LIKES, J. Distribution of Dixon's statistics in the case of an exponential population. *Metrika*, v.16, p.46–54, 1966.

MCMILLAN, R. G. Tests for one or two outliers in normal samples with unknown variance. *Technometrics*, v.13, n.1, p. 87–100, 1971.

MCMILLAN, R. G., DAVID, H. A. Tests for one or two outliers in normal samples with know variance. *Technometrics*, v.13, n.1, p. 75–85, 1971.

MORAN, M. A, MCMILLAN, R. G. Tests for one a two outliers in normal samples with unknown variance: correction *Technometrics*, v. 15, n. 3, p. 637–640, aug. 1973.

MOSTELLER, F., TUKEY, J. W. *Data analysis and linear regression.* Reading, Massachussets : Addison–Wesley, 1977.

MURTHY, R. B. *On tests for outlying observations.* Princeton, New Jersey : Princeton University, 1951. Tese (doutorado em estatística) Princeton University, 1951.

PEARSON , E. S., CHANDRA SEKAR, C. The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, v.28, part. I/II, p. 308–320, 1936.

PEARSON, E. S., STEPHENS, M. A. The ratio of range to standard deviation in the same normal sample. *Biometrika*, v. 51, part. 3/4, p. 484–487, 1964.

PRESCOTT, P. Critical values for a sequential test for many outliers. *Appl. Statis.*, v. 28, n. 1, p. 36–39, 1979.

———— Examination of the behaviour of tests for outliers when more than outlier is present. *Appl. Statis.*, v.27, n.1, p. 10–25, 1978.

QUESENBERRY, C. P., DAVID, H. A. Some tests for outliers. *Biometrika*, v. 48, part. 3/4, p. 379–387, 1961.

SHAPIRO, S. S., WILK, M. B. An analysis of variance test for normality: (complete samples). *Biometrika*, v. 52, part. 3/4, 1965.

SHAPIRO, S. S., WILK, M. B. An Analysis of variance test for the exponential distribution: (complete samples) *Technometrics*, v. 14, n. 2, p. 355–370, 1972.

SHAPIRO, S. S., WILK, M. B. CHEN, M. J. A comparative study of various tests for normality. *JASA*, v. 63, n. 324, p. 1343–1372, 1968.

STEFANSKY, W. Rejeting outliers by maximum normal residual. *Ann. Math. Statist.*, v.42, n.1, p. 35–45, 1971.

STEPHENS, M. A. On the W–test for exponentiality with origin known. *Technometrics*, v. 20, n. 1, p. 33–35, 1978.

TIETJEN, G. L., MOORE, R. H. Some Grubbs–type statistics for the detection of several outliers. *Technometrics*, v. 14, n. 3, p. 583–597, 1972.

WALSH, J. E. *Handbook of non–parametric statistics, II.* Princeton, New Jersey : Van Nostrand, 1965.

A FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE E A PRODUÇÃO DE ESTATÍSTICAS

Lenildo Fernandes Silva*

1 – APRESENTAÇÃO

O presente texto foi elaborado com o objetivo de expor, de forma organizada e sistemática, ainda que de maneira introdutória, os procedimentos técnicos utilizados pelo IBGE, em particular pela Diretoria de Pesquisas, relacionados à concepção de pesquisas e produção de informações estatísticas. As contribuições existentes sobre o assunto estavam dispersas em vários textos, levando-nos a nos dedicar a tarefa de produzir um documento-síntese com vistas a 25ªConferência Internacional de Estatística das Nações Unidas, realizada em fevereiro último em Nova Iorque, conferência à qual o IBGE tem sido convidado e enviado representantes.

Estamos certos de que a elaboração de tal documento se insere no esforço conjunto que esta Diretoria vem fazendo, no sentido de documentar e disseminar o mais amplamente possível suas atividades técnicas.

Finalmente, agradecemos a Pedro Luis do Nascimento Silva, Chefe do Núcleo de Metodologia da Diretoria de Pesquisas do IBGE, pela cuidadosa revisão do texto original e pelos valiosos comentários e sugestões apresentados. Erros e omissões são, no entanto, de nossa exclusiva responsabilidade.

*Diretor de Pesquisas da Fundação Instituto Brasileiro de Geografia e Estatística - IBGE.

2 – INTRODUÇÃO

A Fundação Instituto Brasileiro de Geografia e Estatística — IBGE é uma instituição pública integrante da Administração Federal e subordinada à Secretaria de Planejamento e Coordenação Geral da Presidência da República — SEPLAN/PR, conforme o Estatuto aprovado pelo Decreto número 97.434, de 05/01/89.

Tem por finalidades básicas a pesquisa, a produção, a análise e a difusão de informações e estudos de natureza estatística, geográfica, cartográfica, geodésica, demográfica e sócio-econômica, de recursos naturais e de condições do meio ambiente, necessárias ao conhecimento da realidade física, humana, econômica e social, com vistas, especialmente, à execução de programas e projetos de desenvolvimento nacional.

Para o cumprimento de sua missão institucional o IBGE tem sua estrutura administrativa básica composta de uma Presidência (composta por um Presidente e um Diretor-Geral), uma Diretoria de Pesquisas, uma Diretoria de Geociências, uma Diretoria de Informática, uma Escola Nacional de Ciências Estatísticas e um Centro de Documentação e Disseminação de Informações, além de órgãos de assessoramento superior e de apoio administrativo.

No que respeita especificamente à produção de informações estatísticas, o IBGE tem como atribuição coordenar o Sistema Estatístico Nacional — SEN, ao qual estão vinculadas outras instituições públicas e privadas produtoras de estatísticas.

Como coordenador do SEN, o IBGE tem a seu encargo a orientação, a coordenação e o desenvolvimento, em todo o Território Nacional, das atividades técnicas do Plano Geral de Informações Estatísticas e Geográficas — PGIE (aprovado pelo Decreto número 74.084, de 20/05/74) a cargo de outros órgãos e entidades, além da produção direta de um conjunto de estatísticas econômicas, sociais e demográficas constantes do referido plano.

A definição e revisão periódica do PGIE na sua parte referente à produção de estatísticas resulta de ampla consulta aos mais importantes órgãos produtores e usuários de informações estatísticas, reunidos em assembléia por ocasião da realização das Conferências Nacionais de Estatística — CONFEST.

A produção de informações estatísticas (econômicas, sociais e demográficas) pelo IBGE é de responsabilidade da Diretoria de Pesquisas — DPE que, para tanto, com o apoio da Diretoria de Geociências — DGC no estabelecimento do referencial espacial e na montagem e manutenção da Base Operacional Geográfica (divisão e mapas setoriais), da Diretoria de Informática — DI, para o processamento das informações e administração da Base de Dados, e do Centro de Documentação e Disseminação de Informações — CDDI, no que respeita à editoração e impressão (serviços gráficos) dos resultados das pesquisas e à disseminação dos mesmos através de publicações, tabulações especiais, sistema Telex, terminais de vídeo (*on line*), rede RENPAC, etc. Conta ainda, para o levantamento (coleta) e disseminação das informações estatísticas,

com uma extensa Rede de Unidades Descentralizadas, constituída de 26 (vinte e seis) Delegacias nos estados e territórios e 1 (uma) Delegacia no Distrito Federal, além de cerca de 650 Agências de Coleta, que cobrem todo o Território Nacional.

Com vistas ao cumprimento do seu papel, no que diz respeito à produção de estatísticas, a Diretoria de Pesquisas — DPE do IBGE está organizada sob a forma de Coordenações, Núcleos e Departamentos.

Existem duas coordenações, uma para o Censo Agropecuário — CCA e outra para os Censos Industrial, de Comércio e Serviços — CCE; Núcleos de Planejamento e Supervisão — NPS, de Metodologia — NME e de Documentação — NDO; e Departamentos de Agropecuária — DEAGRO, de Indústria — DEIND, de Comércio e Serviços — DECSE, de Contas Nacionais — DECNA, de Índices de Preços — DESIP, de Emprego e Rendimento — DEREN, de Estatísticas e Indicadores Sociais — DEISO e de População — DEPOP, além de uma Gerência de Suporte Administrativo — GESAD.

Com o objetivo de bem retratar os aspectos mais relevantes da realidade brasileira, o IBGE realiza, através de sua Diretoria de Pesquisas, cerca de 50 pesquisas econômicas, sociais e demográficas. Tais pesquisas, de naturezas as mais diversas — primárias x derivadas, contínuas x eventuais, diretas x indiretas, censitárias x amostrais, etc. — e de periodicidades variadas, abrangem várias áreas temáticas e setoriais.

3 – OS CENSOS

Os Censos são as principais pesquisas realizadas pela DPE, não somente por investigarem, exaustivamente, todo o universo de informantes, embora possam também empregar amostragem na coleta de dados como no caso do Censo Demográfico — ver Metodologia do Censo Demográfico de 1980 (1983) — mas porque propiciam um conjunto abrangente de informações cujo detalhamento espacial e de classificação não pode ser atingido senão numa pesquisa desse tipo, além

de fornecerem a base cadastral indispensável para a realização das pesquisas contínuas (amostrais ou não), além de balizarem algumas importantes questões conceituais.

O IBGE realiza, com periodicidades distintas, os Censos Demográfico e Econômicos (Agropecuário, Industrial, Construção, Comércio, Transportes e Serviços). O Censo Demográfico, de periodicidade mais larga — 10 em 10 anos — investiga características demográficas, econômicas e sociais da população e determinadas características dos domicílios, relativas, especialmente, à infra-estrutura dos mesmos, retratando um quadro estrutural abrangente da sociedade brasileira.

Os Censos Econômicos, de periodicidade também larga — 5 em 5 anos — retratam

a organização da produção de bens e serviços, permitindo, assim, pela sua abrangência, a exemplo do Censo Demográfico, a formação de amplo quadro estrutural da economia brasileira.

Censos	Periodicidades	Executores
Demográfico	10 em 10 anos	DEPOP
Agropecuário	5 em 5 anos	DEPOP
Industrial	5 em 5 anos	DEAGRO/CCA
Industrial	5 em 5 anos	DEIND/CCE
Construção	5 em 5 anos	DEIND/CCE
Comércio	5 em 5 anos	DECSE/CCE
Transportes	5 em 5 anos	DECSE/CCE
Serviços	5 em 5 anos	DECSE/CCE

4 – PESQUISAS CONTÍNUAS

Tendo como importante referencial os Censos, que fornecem a base cadastral e contribuem com a formação da base conceitual e a padronização dos sistemas de classificação (por atividade, ocupação, etc.), realiza-se um conjunto de pesquisas contínuas, de periodicidade mais estreita, que têm por objetivo captar algumas modificações de caráter estrutural (principalmente as pesquisas anuais e semestrais) da economia e da sociedade brasileira, bem como fornecer referencial analítico para o entendimento da conjuntura, ao funcionarem como sinalizadoras das tendências de curto prazo (principalmente as pesquisas mensais ou trimestrais).

Tais pesquisas, resultantes de levantamentos primários, juntamente com os dados de alguns Registros Administrativos, propiciam importantes subsídios aos órgãos de análise e planejamento, bem como à elaboração da síntese econômica (Contas Nacionais) e dos Indicadores Sociais. Elas acham-se distribuídas por áreas temáticas e setoriais, segundo os departamentos responsáveis, conforme indicado a seguir. Para ter acesso a uma descrição resumida de cada uma das pesquisas mencionadas, inclusive dos Censos, consultar (Principais Características das Pesquisas Econômicas, Sociais e Demográficas, 1988).

4.1 – Econômicas

4.1.1 — Departamento de Agropecuária — DEAGRO

Anuais

- Pesquisa Agrícola Municipal — PAM
- Pesquisa Pecuária Municipal — PPM
- Pesquisa Extrativa Vegetal e Silvicultura — PEVS

Semestrais

- Estatística da Pesca — PESCA
- Pesquisa de Estoques — ESTOQ

Trimestrais

- Pesquisa de Ovos de Galinha — POG

Mensais

- Levantamento Sistemático da Produção Agrícola — LSPA
- Pesquisa Mensal de Abate de Animais — ABATE
- Pesquisa Mensal de Leite — LEITE

4.1.2 – Departamento de Indústria — DEIND

Anuais

- Pesquisa Industrial Anual — PIA

Mensais

- Pesquisa Industrial Mensal/Produção Física — PIM/PF
- Pesquisa Industrial Mensal/Dados Gerais — PIM/DG
- Nova Pesquisa da Construção — NOPEC

4.1.3 – Departamento de Comércio e Serviços — DECSE

Anuais

- Pesquisa de Transporte Rodoviário — PATR
- Pesquisa Anual de Comércio — PAC
- Pesquisa de Meios de Hospedagem — PMH

Trimestrais

- Relatório de Indicadores do Setor Serviços

Mensais

- Pesquisa Mensal de Comércio — PMC

4.1.4 – Departamento de Contas Nacionais — DECNA

Anuais

- Contas Nacionais Consolidadas
- Novo Sistema de Contas Nacionais

- Matriz de Insumo-Produto
- Produto Interno Bruto Real — PIB Real
- Estatísticas Econômicas das Administrações e Empresas Públicas

Trimestrais

- Produto Interno Bruto Real (PIB Real) Trimestral

4.1.5 – Departamento de Índices de Preços — DESIP

Qüinqüenais

- Pesquisa de Orçamentos Familiares — POF
- Pesquisa de Especificação de Produtos e Serviços — PEPS
- Pesquisa de Locais de Compra — PLC
- Pesquisa de Especificação de Materiais

Mensais

- Sistema Nacional de Índices de Preços ao Consumidor — SNIPC
- Sistema Nacional de Pesquisa de Custos e Índices da Construção Civil — SINAPI

4.2 – Sociais

4.2.1 – Departamento de Emprego e Rendimento — DEREN

Anuais

- Pesquisa Nacional por Amostra de Domicílios — PNAD — Corpo Básico
- Pesquisa Nacional por Amostra de Domicílios — PNAD — Suplemento Especial (juntamente com outros departamentos da DPE)

Mensais

- Pesquisa Mensal de Emprego — PME

4.2.2 – Departamento de Estatísticas e Indicadores Sociais — DEISO

Trienal

- Relatório de Indicadores Sociais

Anuais

- Relatório de Indicadores Sociais — Sistema Contínuo
- Assistência Médico Sanitária — AMS
- Pesquisa Sindical
- Perfil Estatístico de Crianças e Mães
- Pesquisas de Cultura
- Pesquisas do Ministério da Justiça
- Pesquisas de Educação e Desporto

4.3 – Demográficas

4.3.1 – Departamento de População — DEPOP

Trimestrais

- Estatísticas do Registro Civil

Sem Periodicidade Definida

- Estudo das Migrações Internas
- Estimativas Rotineiras de População
- Estudos de Mortalidade e Fertilidade

Dentre tais pesquisas, cabe um esclarecimento especial em relação à Pesquisa Nacional por Amostra de Domicílios — PNAD. Embora esteja sob a responsabilidade do Departamento de Emprego e Rendimento — DEREN, a PNAD caracteriza-se como uma pesquisa de múltiplos propósitos, vez que, além das informações sobre rendimento e emprego, fornece, nos anos intercensitários, estimadores para as variáveis demográficas, informações para a construção dos Indicadores Sociais (Corpo Básico), bem como permite a exploração mais detalhada, a cada ano, de temas que tenham afinidade com o seu conteúdo básico, através da adição de Suplementos Especiais (veja Metodologia da Pesquisa Nacional por Amostra de Domicílios na década de 70, 1981). Vale notar que, presentemente, o DEREN tem realizado reflexões com vistas a fixar sua identidade na área de estatísticas do trabalho.

Todas as pesquisas anteriormente enumeradas e realizadas diretamente pelo IBGE procuram assegurar a produção de informações estatísticas que constam do Plano Geral de Informações Estatísticas — PGIE. Desse plano faz parte, ainda, um outro conjunto de informações, também de responsabilidade do IBGE, mas produzidas por pesquisas e levantamentos a cargo de outras instituições, sob a forma de delegação.

Entre outros motivos, isso implica que a Coordenação pelo IBGE do Sistema Estatístico Nacional — SEN, com vistas à execução do PGIE, deve ser exercida de maneira a garantir a qualidade e homogeneidade das estatísticas produzidas. Para tanto, o IBGE tem procurado aumentar o relacionamento com os demais órgãos produtores de estatísticas, buscando maior articulação com vistas à integração de suas atividades.

5 – A PRODUÇÃO DE ESTATÍSTICAS

A decisão de se realizar uma nova pesquisa — seja em atendimento a uma reflexão interna ou a uma demanda externa — está referida e condicionada, inicialmente, às prioridades definidas no Plano Geral de Informações Estatísticas — PGIE e sempre se dá em atendimento à necessidade de informações básicas, à elaboração da Síntese Econômica e dos Indicadores Sociais. Tal decisão visa geralmente a propiciar informações

novas e mais qualificadas para subsidiar a análise e o planejamento, ou o preenchimento de lacunas acerca do conhecimento de certas áreas temáticas e setoriais.

As informações a serem geradas podem ser de natureza estrutural e/ou conjuntural, desde que relevantes para o entendimento da economia e da sociedade brasileira. Dessa forma, o elenco de pesquisas realizadas pelo IBGE resulta, necessariamente, de um processo dinâmico, que exige o permanente exercício da reflexão sobre a realidade — em permanente mutação — objeto de estudo.

Nesse sentido, o Programa de Trabalho Anual — PTA (ver, por exemplo, o Plano de Trabalho Anual de 1989 da Diretoria de Pesquisas em Plano de Trabalho Anual de 1989, da Diretoria de Pesquisas, contempla a existência de projetos tanto de produção quanto de revisão de pesquisas — todos com recursos específicos — dada a relevância e complexidade de cada uma dessas etapas do processo de pesquisa.

Uma vez tenha sido definido, em linhas gerais, o objeto da pesquisa, é montada uma equipe básica de trabalho. Tal equipe, freqüentemente de caráter multidisciplinar, envolve profissionais de formações distintas, tais como estatísticos, economistas, sociólogos, agrônomos, analistas de sistemas, programadores, etc.

Nesse aspecto, vale destacar que a formação de quadros com experiência na produção de estatísticas demanda tempo e se complementa no processo de trabalho, dependendo a preservação de tais quadros de adequadas condições financeiras e materiais de trabalho, nem sempre presentes nas instituições de pesquisas.

Esse problema, associado às dificuldades de renovação — oxigenação — do quadro técnico, constitui foco de constantes dificuldades.

O planejamento de uma pesquisa é uma atividade complexa que envolve questões de ordem teórica, conceitual, temporal, de recursos humanos, financeiros, etc. A sua administração se dá, na maioria das vezes, com recursos escassos.

A implantação de uma nova pesquisa no âmbito da DPE é precedida de discussões internas ao Corpo Técnico da instituição.

Tais discussões contam, na maioria das vezes, com o concurso — em determinadas etapas do processo — de consultores pertencentes aos quadros de outros órgãos produtores e/ou usuários de informações estatísticas (em particular, de Universidades e Institutos de Pesquisas) com tradição e experiência na área objeto de estudo.

Dependendo do objeto ou tema a ser investigado, e da natureza da pesquisa, importantes decisões iniciais devem ser tomadas com respeito ao âmbito da investigação, abrangência espacial, periodicidade (aí incluído o Calendário de Coleta), unidade de investigação, método de coleta das informações, questionário e desenho da pesquisa.

A definição do âmbito da investigação está intimamente ligada à identificação do *locus* de ocorrência mais significativa e relevante do fenômeno a ser estudado. A definição do âmbito determina o nível (limite) ao qual se circunscreve o objeto a ser investigado. Juntamente com a definição da unidade de investigação e da abrangência espacial, a definição do âmbito identifica a população alvo da pesquisa.

Quanto à abrangência espacial, depende da extensão de ocorrência do fenômeno, de sua representatividade espacial, de sua relevância para o entendimento de questões regionais, da disponibilidade de recursos, etc. Em países como o Brasil, de dimensão continental, são elevados os custos para que possa ser feita uma cobertura ampla, ao mesmo tempo em que a diversidade física, econômica e social exige representações particulares de todo o espaço.

Tais situações, excluídas as investigações censitárias, levam algumas vezes à construção de amostras de pouca representatividade regional, preparadas apenas para dar resultados a nível nacional, e outras vezes à necessidade de utilização de cortes no espaço a ser investigado, como é o caso das pesquisas por regiões metropolitanas selecionadas. Dessa forma, cada situação particular determinará o recorte e a abrangência espacial requeridos para representar o objeto em estudo.

A periodicidade, a exemplo de questões como âmbito e abrangência espacial, depende de um conjunto de fatores, nem sempre convergentes, relacionados à freqüência e duração de ocorrência do evento, sua estabilidade/estacionalidade, existência de registro, sua relevância para decisões de curto prazo, etc.

Está, também, condicionada pela disponibilidade de recursos frente à relevância do evento a ser registrado.

Quanto à época de coleta, deve tomar em conta a ocorrência temporal do fenômeno, a existência de registros sobre os mesmos e o calendário (distribuição temporal e espacial da coleta) das demais pesquisas existentes, devido ao seu impacto na utilização de uma infra-estrutura de coleta de múltiplo atendimento, como é a Rede de Coleta do IBGE.

Dependendo da natureza do objeto ou tema a ser pesquisado, as pesquisas podem ter como unidades de investigação estabelecimentos ou empresas (econômicas), ou domicílios, famílias e pessoas (sócio-econômicas e demográficas). Em ambos os casos, o levantamento pode ser exaustivo, isto é, de caráter censitário (quando investigar todas as unidades da população alvo), ou parcial, quando apoiado na utilização de amostras (probabilísticas ou intencionais) de unidades da população alvo da pesquisa.

As pesquisas amostrais são de particular importância por permitirem ganhos de tempo e redução de custos sem perda de qualidade dos resultados, embora na dependência do desenho e tamanho da amostra e das características das distribuições das variáveis de interesse na população alvo.

Por outro lado, a existência de bons cadastros derivados das pesquisas censitárias e/ou de outras fontes (registros administrativos tais como o Imposto de Renda, RAIS, etc.), permanentemente atualizados, é de importância fundamental para assegurar a possibilidade de desenhar boas amostras, sem o que não é possível esperar boa qualidade dos resultados obtidos.

Todas essas decisões (âmbito, abrangência, periodicidade e unidade de investigação) estão condicionadas, e também condicionam, a definição do conteúdo e do desenho da

pesquisa, isto é, a sua expressão sob a forma de quesitos de investigação, o próprio desenho do questionário e a escolha do método de cobertura da população alvo (censitária ou amostral).

Este é um momento crucial na concepção de uma pesquisa, já que a formulação de questões, o seu encadeamento, as alternativas de respostas, os quesitos abertos, o catálogo de escritores, etc., podem introduzir erros (tais como questões incorretas ou mal formuladas, indução de respostas, dificuldade para a realização das entrevistas, etc.), que podem tornar inúteis ou pouco representativos os resultados obtidos. Por outro lado, o desenho do questionário é fator importante na preservação da lógica interna à investigação, na garantia de facilidades para a realização de testes de qualidade do preenchimento e de consistência interna, bem como para facilitar a codificação e digitação dos dados, permitindo redução de custos e da margem de erros.

Antes de ser executada ou implantada, toda pesquisa é testada em caráter preliminar ou experimental, de maneira a se poder avaliar se o objeto de estudo está sendo corretamente investigado, se o questionário é consistente, se o tempo previsto para coleta é suficiente, etc. O teste de campo, isto é, a realização de uma etapa experimental, é fundamental para o sucesso de uma pesquisa.

Em se tratando de pesquisa por amostra, é de fundamental importância a etapa de definição do plano amostral (métodos para seleção e expansão da amostra, e definição de seu tamanho). A construção da amostra depende ainda da existência de bons cadastros (de áreas, domicílios, empresas, estabelecimentos) como já mencionado.

Em seguida é executada, no todo ou em parte, a seleção efetiva das unidades que vão compor a amostra da pesquisa. Em alguns casos de pesquisas que utilizem amostras de áreas (como a PNAD e a PME, por exemplo) a seleção final das unidades a serem pesquisadas será feita em campo, durante a coleta da pesquisa (ver, a esse respeito, Metodologia da Pesquisa Nacional por Amostra de Domicílios na década de 70, (1981) e Metodologia da Pesquisa Mensal de Emprego de 1980 (1983a).

Cada um dos Departamentos vinculados à Diretoria de Pesquisas tem, para a execução de tarefas ligadas ao planejamento amostral, um grupo de apoio metodológico, do qual fazem parte estatísticos.

A Diretoria mantém ainda em sua estrutura um Núcleo de Metodologia — NME, composto especialmente por estatísticos amostristas, que também prestam apoio aos Departamentos. Tal Núcleo, no entanto, tem como principal meta trabalhar na "fronteira", isto é, ser a vanguarda das transformações metodológicas na Diretoria de Pesquisas.

Após definidos o questionário (conteúdo e desenho) e o desenho da pesquisa, uma série de atividades complementares é desencadeada simultaneamente.

Uma das atividades mais importantes está relacionada com o treinamento das equipes de instrutores, supervisores e recenseadores/entrevistadores. Para tanto, faz-se necessária a elaboração de um Plano de Treinamento, de Manuais de Instrução, e

de outros instrumentos de apoio à coleta, capazes de permitir o esclarecimento de questões complexas, de estabelecer normas e homogeneizar procedimentos de coleta, de maneira que não ocorram interpretações diferentes para situações comuns. Recentemente, começou-se a utilizar de modernos recursos audiovisuais (vídeo-teipe) como ferramenta auxiliar no treinamento dessas equipes, como ocorreu no treinamento dos recenseadores que trabalharam na coleta do Censo Demográfico Experimental de Limeira em 1988.

Uma vez tenha sido testado e aprovado o questionário, a pesquisa vai a campo em sua versão definitiva. A coleta é realizada pelos agentes lotados nas Agências de Coleta das Delegacias Estaduais do IBGE, os quais recebem os questionários e suas etiquetas de identificação e/ou orientação específica, além de serem treinados para cada coleta específica da qual irão tomar parte.

Devido aos custos elevados, ainda são poucas as pesquisas nas quais usamos identificação prévia, através de formulários contínuos, como já fazemos no SNIPC, no SINAPI, na PIA, na PIM-PF e na PIM-DG. Algumas pesquisas, devido à periodicidade, intensidade e/ou complexidade do levantamento, possuem equipe de coleta permanente ou exclusiva, como é o caso do SNIPC e da PME.

Em se tratando de coleta, cabe esclarecer que o IBGE usa, para a quase totalidade de suas pesquisas, procedimentos tradicionais de entrevista direta, quando muito deixando questionários com um conjunto de grandes empresas selecionadas, para posterior recolhimento. ^ão existe coleta (envio e recebimento) por via postal, nem informantes voluntários que prestem informações periódica e permanentemente.

Tampouco são utilizados meios eletrônicos ou magnéticos de coleta e recebimento de informações tais como fitas magnéticas, disquetes, entrevistas por telefone. Apesar disso, já foi testada, mas não implantada devido ao grande investimento requerido, a utilização de microcoletores de dados (microcomputadores que simulam questionários e que registram as respostas digitadas pelos entrevistadores em cartuchos magnéticos), a qual foi experimentada na coleta de preços do SNIPC com relativo sucesso. (ver, a respeito, Principais Características das Pesquisas Econômicas, Sociais e Demográficas (1988 b).

Apesar de não estar sendo usado, o sistema de coleta usando microcoletores teria a vantagem de eliminar completamente etapas tais como empastamento e digitação de questionários, além de permitir que críticas básicas de preenchimento (existência) e mesmo de consistência das informações coletadas fossem realizadas ainda durante a coleta, garantindo melhoria da qualidade dos dados e diminuindo o trabalho e o custo de verificações posteriores. Seu inconveniente é que para ser aplicado com boa margem de segurança exigiria que as unidades a serem pesquisadas estivessem pré-identificadas, ou então que informações alfanuméricas (nomes, etc.) tivessem que ser digitadas pelos entrevistadores.

O trabalho de coleta, tanto no caso das pesquisas domiciliares quanto das pesquisas

por estabelecimentos, depende da existência de uma boa Base Operacional Geográfica, a qual precisa ser constantemente atualizada, de maneira a fornecer mapas por setores adequados, atualizados e bem ilustrados.

O trabalho de atualização da Base Operacional também leva em conta a necessidade de permitir a recuperação de informações por áreas mínimas de investigação, de maneira que as mesmas possam ser comparáveis historicamente (série histórica preservada). Tal trabalho é atualmente realizado pela Gerência de Base Operacional do Departamento de Cartografia da Diretoria de Geociências que conta com a participação dos Setores de Base Operacional existentes em todas as Delegacias Estaduais do IBGE.

Todo o relacionamento operacional da Diretoria de Pesquisas com a Rede de Coleta (Delegacias e Agências) é supervisionado pelo Núcleo de Planejamento e Supervisão — NPS da DPE, que acompanha a execução de todo o Programa de Trabalho Anual — PTA, bem como define conjuntamente o Calendário de Coleta de todas as pesquisas, estipulando prazos e determinando as datas limites, além de administrar a alocação dos recursos humanos e financeiros necessários à sua efetivação.

Encerrada a coleta, os questionários são enviados aos Departamentos específicos da Diretoria de Pesquisas, no Rio de Janeiro, com vistas ao início da etapa de processamento que, para a maioria das pesquisas, se faz centralizadamente.

Devido às dificuldades (volume/tempo) de processamento centralizado, o IBGE tem equipado algumas de suas principais Delegacias Estaduais, bem como aquelas localizadas estrategicamente, do ponto de vista regional, para a realização de algumas etapas do processamento de forma descentralizada. Tal descentralização inclui basicamente as tarefas de empastamento, controle de recepção, transcrição e crítica quantitativa dos dados, e não inclui as etapas de crítica qualitativa, imputação, expansão e tabulação. Essa descentralização tem permitido a agilização dos procedimentos e uma redução nos prazos de apuração.

Outra atividade diz respeito ao desenvolvimento dos sistemas de apuração da pesquisa, aí incluídos os sistemas de empastamento e/ou controle de recepção de questionários, entrada de dados (digitação), codificação, crítica e correção (ou acerto), imputação automática, expansão e tabulação.

A concepção de tais sistemas requer a participação, sob coordenação da área fim (DPE) — aquela que concebe o conteúdo da pesquisa — de analistas de sistemas e programadores da própria DPE e da Diretoria de Informática — DI, com vistas ao desenvolvimento dos referidos sistemas e programas.

Para a transcrição ou entrada dos dados ainda utilizamos o sistema de digitação em máquina (terminais de vídeo acoplados a um minicomputador que armazena os dados em fitas magnéticas), processo que consome muito tempo e utiliza grande número de digitadores (preparadores de dados), trabalhando em vários turnos diários.

Recentemente, avaliamos que deveríamos aguardar mais algum tempo para intro-

duzirmos o processo de leitura ótica (ver, a respeito, Silva (1988)), devido ao seu alto custo e à rápida evolução tecnológica do setor, que tem levado à inexistência de processos e equipamentos estabilizados. A introdução de tal processo, no entanto, certamente contribuirá para eliminar os inconvenientes de alto custo e espaço físico requerido para a armazenagem dos questionários, além daqueles relacionados à administração de pessoal temporário, e possivelmente dispensará a etapa de empastamento. Sem dúvida, não se poderá esquecer da atraente alternativa oferecida pelo emprego de microcoletores, conforme já foi mencionado.

Quanto à codificação, estamos evoluindo para o sistema de codificação assistida por computador, que consiste na construção de arquivos de descritores os mais completos possíveis, para as mais variadas situações características de quesitos abertos (ocupação, religião, municípios, educação, produtos, etc.), a serem consultados *on line* — via um sistema de recuperação que pesquisa o catálogo de descritores a partir da entrada pelo operador de uma ou mais palavras-chaves contidas nas descrições dos itens a serem codificados.

A consulta *on line* ao catálogo de descritores tem permitido maior rapidez e segurança quanto ao resultado do processo de codificação nas pesquisas em que esse sistema vem sendo utilizado, como por exemplo na codificação de produtos do Censo Industrial de 1985 — (ver & Silva e outros(1987)).

Mais recentemente, o sistema de codificação assistida por computador evoluiu no sentido de uma maior automação do processo, a partir da idéia de digitação dos próprios textos dos descritores e da codificação automática dos mesmos por um programa de codificação executado em *batch*. Esse sistema já foi empregado em caráter experimental para codificar os dados de religião, migração, último grau concluído, ocupação e atividade investigados no Censo Demográfico Experimental de Limeira de 1988, com relativo sucesso, conforme reportam Mansoldo(1989) e Silva(1989).

No que respeita à etapa de crítica, um primeiro procedimento está relacionado à realização de uma pré-crítica (crítica de preenchimento) que dá origem aos primeiros Boletins de Consulta às Delegacias (quando realizada na sede), muito usados durante todo o processo de crítica. Além disso, atenção especial é dada à definição dos "planos de crítica", conjuntos de regras para verificação da existência (Crítica Quantitativa) e consistência (Crítica Qualitativa) das informações. Planos de Crítica bem definidos facilitam o desenvolvimento dos sistemas de apuração e tornam mais rápido e eficiente o processamento das informações, assegurando um padrão de qualidade para os dados.

Algumas de nossas principais pesquisas já estão utilizando sistemas de crítica *on line*, o que permite a realização da crítica e dos acertos simultaneamente (em tempo real), eliminando assim a necessidade de impressão de várias listagens de erros (processamento via *batch* tradicional) correspondentes às diversas passagens de crítica, até que os dados sejam aceitos.

Este processo permite redução de tempo e de custos e maior qualidade na apuração

das informações. Para tanto, todos os departamentos contam com uma bateria de terminais de vídeo (em expansão) e microcomputadores interligados com o computador central.

Além disso, tem sido desenvolvido um esforço conjunto com a Diretoria de Informática no sentido de passar a utilizar uma metodologia mais moderna e eficaz para integrar as tarefas de crítica, correção e imputação das informações, para o que está em desenvolvimento um pacote de Crítica, Imputação e Tabulação (**Cripta**) — ver Duarte, Barbosa e Hanono(1988).

Uma vez concluídas todas as etapas anteriores, segue-se o desenvolvimento dos procedimentos de imputação automática (no caso de pesquisas que adotem essa forma de correção dos dados) e expansão (no caso de pesquisas por amostragem).

A imputação automática é uma modalidade de correção de dados adotada desde 1980 em nossos Censos, e desde então em algumas de nossas pesquisas contínuas (como a PNAD e a PME, por exemplo). Sua finalidade é preencher eventuais lacunas das informações individuais de algumas unidades pesquisadas que não tenham sido obtidas até o momento imediatamente anterior à expansão, bem como substituir por valores plausíveis alguns valores originalmente respondidos que porventura tenham sido rejeitados como inconsistentes ou inválidos pelo sistema de crítica, sem que tenha sido possível (por razões de custo ou outras) recorrer ao informante original para solucionar o problema.

No Censo Demográfico e na PME a metodologia usada para imputação combina regras de imputação determinísticas com aplicação do método *hot-deck* seqüencial — ver Metodologia do Censo Demográfico de 1980...(1983). No Censo Agropecuário utiliza-se imputação por médias e nos Censos Econômicos adotam-se diversos tipos de imputação determinística orientada por especialistas temáticos.

Uma vez encerrada esta etapa, em se tratando de uma pesquisa amostral, executa-se o trabalho de expansão da amostra, empregando para isso estimadores que permitirão a obtenção de resultados válidos para o conjunto da população alvo da pesquisa. Em função dos problemas operacionais (volume de questionários) e de custos, o IBGE tem procurado trabalhar com amostras cada vez menores, embora mantendo um determinado nível mínimo de confiabilidade e fidedignidade para as estimativas produzidas, de modo que estas bem representem o fenômeno objeto de estudo.

Isto tem exigido esforços no sentido de redesenhar e reduzir algumas amostras de pesquisas já tradicionais, como o Censo Demográfico (ver Carvalho e outros(1988)) e Pinheiro e Lima(1989), a PNAD (ver Menezes e outros(1987)) e a PME (ver Silva e Moura(1989), bem como a introdução do método de amostragem em outras pesquisas nas quais isso não era feito (na PIA, por exemplo, desde 1981 (ver Pesquisa Industrial 1982-1984 (1988a)).

Um outro aspecto já tradicional no tratamento da divulgação de estatísticas elaboradas a partir de pesquisas por amostragem é a publicação de medidas de precisão

(coeficientes de variação — CVs) das estimativas, permitindo que os usuários avaliem com facilidade o grau de confiabilidade e fidedignidade dos resultados que a amostra permitiu obter. Em algumas pesquisas, como o Censo Demográfico de 80 e a PNAD, empregam-se as chamadas Generalized Variance Functions — Wolter(1985) — para indicar o nível de precisão das estimativas sem ter que recorrer à publicação do Coeficiente de Variação individual de cada uma delas.

Nessa mesma linha, o Censo Demográfico também incorpora estudos de avaliação de erros não amostrais que dão origem a retificações nas estimativas de população publicadas a partir da apuração dos dados básicos. Essa prática ainda não pode ser estendida a outras pesquisas devido às dificuldades técnicas e de recursos para sua realização em moldes satisfatórios.

Uma das atividades mais importantes executada pela equipe responsável por um novo projeto de pesquisa diz respeito à definição do Plano Tabular. Tal plano deve contemplar a divulgação por meio de tabelas dos resultados mais significativos, atendidas as mais relevantes possibilidades de cruzamentos de informações disponíveis na pesquisa. A sua elaboração é um trabalho conjunto dos técnicos do Departamento temático específico da Diretoria de Pesquisas alocados ao projeto, apoiados por Analistas de Sistemas e Programadores da Diretoria de Informática que ficarão responsáveis pela elaboração dos programas necessários para a produção das tabelas desejadas, geralmente empregando os sistemas *Atlas* e *Prometeu* desenvolvidos pelo IBGE para essa finalidade.

Uma vez pronto o Plano Tabular, os resultados disponíveis são imediatamente liberados para a Base de Dados administrada pela Diretoria de Informática, onde os mesmos passarão a ficar disponíveis para consulta, guardadas certas restrições (sigilo).

Por isso mesmo a construção do Plano Tabular deve levar em conta certas restrições para impedir a identificação de informações individualizadas a partir dos dados divulgados (desidentificação para manutenção do sigilo), para minimizar a divulgação de quesitos com freqüências muito baixas, para evitar a construção de tabelas contendo estimativas cujos coeficientes de variação (CVs) sejam muito elevados (o que estaria indicando ser a amostra pouco adequada para o fornecimento dessas estimativas), bem como deve também levar em conta o nível mínimo de desagregação que a amostra suporte em termos de divulgação.

Além do Plano Tabular tradicional já empregamos várias outras formas de divulgação, todas elas de responsabilidade do Centro de Documentação e Disseminação de Informações — CDDI. A primeira delas diz respeito àquela parte do plano (conjunto de tabelas) que será publicada em volumes de divulgação dos resultados, visto serem as publicações, pelo seu baixo custo, a forma de mais ampla consulta aos resultados das pesquisas. Uma outra parte do plano (conjunto de tabelas), embora não publicadas, estará disponível para "pronta entrega" mediante solicitação ao IBGE.

Além dessas formas de acesso aos resultados das pesquisas, os usuários poderão uti-

lizar o sistema Telex de acesso ao nosso Banco de Dados; Salas de Consulta instaladas nas Delegacias, equipadas com terminais de vídeo; interligar computadores ao nosso Banco de Dados através da RENPAC, vídeo-texto, bem como solicitar a elaboração de listagens especiais, comprar fitas magnéticas (ou disquetes) com dados desidentificados e/ou relativos a amostras reduzidas de nossas pesquisas.

Todo o programa editorial da Diretoria de Pesquisas, tanto no que diz respeito à impressão de questionários, quanto à publicação de resultados das pesquisas, é coordenado pelo Núcleo de Documentação — NDO.

Além dessas atividades, o NDO é também responsável pela edição da revista mensal "Indicadores IBGE", que divulga o resultado de nossas principais pesquisas conjunturais; pela preparação do material produzido pela DPE para o *Anuário Estatístico do Brasil*; pela Biblioteca Setorial; pela organização de todos os documentos de trabalho produzidos na DPE; pela Série "Relatórios Metodológicos" que documenta as metodologias de todas as nossas pesquisas e pela publicação dos "Textos para Discussão", que registram a parte mais significativa da produção do Corpo Técnico da DPE, relativa à sistematização, proposição metodológica e análise dos resultados de nossas pesquisas.

Em caso de necessidade de maiores esclarecimentos, ou de se tratar de usuários cujas demandas não puderam ser satisfeitas através da consulta às nossas publicações ou à nossa Base de Dados.[1]

[1] Núcleo de Atendimento ao Usuário — CDDI/N — Rua Gen. Canabarro, 666, 1º andar, Maracanã — [Rio de Janeiro - Tel.: (021) 234-2043, r. 281].

BIBLIOGRAFIA

CARVALHO, J.H.S. et al. *Recomendações para o censo demográfico experimental.* Rio de Janeiro, 1988. 30 p. (inédito).

DUARTE, M.A.G., BARBOSA, D.M.R., HANONO, R.M. *Sistema CRIPTA.* Rio de Janeiro, 1988. 30 p. (inédito).

MENEZES, A.C.F. et al. *Pesquisa nacional por amostra de domicílios: redução do número de domicílios na amostra e propostas de novos estimadores.* Rio de Janeiro, 1987. 256 p. (inédito).

METODOLOGIA da Pesquisa Nacional por Amostra de Domicílios na década de 70. Rio de Janeiro : IBGE, 1981. 698 p. (Série Relatórios Metodológicos/IBGE; v. 1).

METODOLOGIA da Pesquisa Mensal de Emprego de 1980. Rio de Janeiro : IBGE, 1983. 89 p. (Série relatórios metodológicos/IBGE; v. 2).

METODOLOGIA do Censo Demográfico de 1980. Rio de Janeiro : IBGE, 1983. 477 p. (Série Relatórios Metodológicos/IBGE; v. 4).

PESQUISA INDUSTRIAL: dados gerais, 1982-84. Brasil, grandes regiões e unidades da federação. Rio de Janeiro : IBGE, v. 9, 1988.

PINHEIRO, J.C.R.C., LIMA, J.M. *Avaliação dos efeitos da redução da fração amostral do censo demográfico de 90.* Rio de Janeiro : Escola Nacional de Ciências Estatísticas, 1988. 48 p. (Relatórios técnicos/Escola Nacional de Ciências Estatísticas; 4/88).

MANSOLDO, H.F. Brasil - Censo demográfico experimental. In: SEMINÁRIO LATINOAMERICANO E CARIBE, 1989, Rio de Janeiro. *Captura e limpieza de datos estadisticos.* Rio de Janeiro : CEPAL, IBGE, 1989. 10 p.

IBGE. Diretoria de Pesquisas e Inquéritos. *Principais características das pesquisas econômicas, sociais e demográficas* - DPE/IBGE. Rio de Janeiro : IBGE, 1988. 154 p. (Textos para discussão/IBGE, DPE; v. 1, n. especial).

PROJETO aquisição de dados - subprojeto pesquisas contínuas - viabilidade técnico financeira - resumo gerencial. Rio de Janeiro : IBGE, 1988. 362 p. (inédito).

SILVA, A.C.M. Codificação do questionário de amostra do censo demográfico experimental de Limeira. In: SEMINÁRIO LATINOAMERICANO E CARIBE, 1989, Rio de Janeiro. *Captura e limpieza de datos estadisticos.* Rio de Janeiro : CEPAL, IBGE, 1989. 14 p.

SILVA, P.L.N. *Relatórios de viagens aos Estados Unidos no período de 12-25 de março de 1988.* Rio de Janeiro : IBGE, 1988. 406 p.

RESUMO

Este artigo foi escrito com vistas à divulgação do atual estágio do processo de produção de estatísticas no IBGE, por ocasião da 25ª Conferência Internacional de Estatística promovida pelo "Statistical Office" das Nações Unidas e realizada em fevereiro de 1989, em Nova Iorque. O artigo descreve aspectos ligados à missão institucional do IBGE, sua organização interna, relaciona suas principais pesquisas e censos, e descreve, em linhas gerais, as principais etapas do processo de produção de informações estatísticas, mencionando e destacando experiências recentes com uso de metodologias avançadas em algumas pesquisas.

ABSTRACT

This paper was written to describe the "state of the art" of the process of producing statistics at IBGE. It was presented at the 25th International Statistical Conference, promoted by the United Nations Statistical Office, held at New York in February, 1989. The paper describes aspects related to IBGE primary objectives, its internal organization, lists the major censuses and surveys and describes, in general guidelines, the main steps in the process of producing statistical information, mentioning and remarking on the recent experimentation with advanced methodologies in some surveys.

USO DE AMOSTRAGEM EM SIMULAÇÃO DE LEGISLAÇÃO TRIBUTÁRIA

José Carlos da R. C. Pinheiro*
e
Manuel Martins Filho**

1 – O PORQUÊ DA AMOSTRA DO IRPF

A cada ano, o Ministério da Fazenda, através da Secretaria da Receita Federal – SRF, altera alguns dos parâmetros que definem a Legislação do Imposto de Renda Pessoa Física – IRPF.

Tais modificações pretendem alcançar metas previamente estabelecidas pelo governo, tais como: aumentar a arrecadação proveniente do tributo (IRPF), promover uma melhor distribuição de renda entre a população, alterar a estrutura da declaração no sentido da simplificação, estimular alguns tipos de investimentos, etc.

Alterações na legislação do IRPF geram, portanto, a necessidade de se estimar previamente os efeitos delas decorrentes, uma vez que a adoção de uma nova legislação promoverá reflexos do ponto de vista econômico, fiscal e administrativo.

Assim, por exemplo, uma correção do Imposto Retido na Fonte em um percentual acima da taxa de inflação verificada no ano-base promoverá um aumento no número de contribuintes com direito a restituição. Da mesma forma, a tributação de rendimentos

*Professor de ensino superior da Escola Nacional de Ciências Estatísticas – ENCE.
**Técnico do SERPRO.

considerados hoje não-tributáveis promoverá um aumento de arrecadação, além de ter um efeito redistributivo importante sobre o perfil de renda da população.

No primeiro exemplo poderíamos estar interessados em avaliar não somente o aumento do número de contribuintes com direito a restituição, mas também em avaliar de quanto aumentará o total de Imposto a Restituir.

No segundo exemplo, além de quantificarmos o aumento de arrecadação promovido pela tributação de rendimentos não-tributáveis, poderíamos avaliar, também, o efeito redistributivo da adoção desta medida, medindo, por exemplo, a variação da alíquota efetiva por classe de renda total.

Questões dessa natureza fazem parte do conjunto de problemas que são colocados à Secretaria da Receita Federal –SRF. Um mecanismo que permita medir, de forma rápida e flexível, os efeitos (na população de declarantes) de alterações na legislação do IRPF é extremamente útil para a autoridade tributária encarregada de gerir este tributo.

Atendendo a uma solicitação da SRF, o SERPRO desenvolveu o Sistema Simulador de Legislações, que vem sendo usado ao longo dos últimos 12 anos como uma ferramenta importante no apoio à tomada de decisão na área tributária.

O Sistema Simulador é basicamente um programa de computador, que tem como entrada a legislação que se pretende avaliar, e uma amostra estratificada do universo de declarantes do IRPF. Para cada declarante da amostra, a declaração é refeita à luz das novas regras que estão sendo avaliadas (nova legislação).

O sistema gera como saída um conjunto de indicadores, tais como: arrecadação total, total de imposto a pagar, total de imposto a restituir, alíquotas efetivas por classe de renda, etc.

Dado o exposto, colocam-se duas questões relevantes. A primeira delas é relativa ao programa, que deverá ser suficientemente flexível para aceitar qualquer tipo de alteração que se deseje fazer na legislação, e ao mesmo tempo ser capaz de gerar, de forma simples e rápida qualquer tipo de relatório, tabela, etc.

A segunda questão nos remete ao lado estatístico do Sistema Simulador, que é o dimensionamento, coleta e uso efetivo de amostras estratificadas. O uso de amostragem neste caso é fundamental, dado que a população de declarantes do IRPF (exercício-1986) é composta por aproximadamente 8 600 000 declarantes.

No capítulo seguinte será analisado, resumidamente, o histórico da experiência do SERPRO na área de amostragem.

No Capítulo 3 será discutida a metodologia adotada na obtenção da amostra. Na Seção 3.1 será apresentado o problema da estratificação. As escolhas da variável de estratificação e do número de estratos serão vistas com detalhes.

Na Seção 3.2 o ponto central será a alocação da amostra. Além da metodologia adotada para a alocação, serão discutidos os seguintes pontos: escolha das variáveis-alvo, a alocação propriamente dita e os ganhos estimados com a estratificação.

A seleção da amostra será abordada na Seção 3.3. A parte operacional será discutida a partir do funcionamento do Programa Amostrador, bem como do uso do gerador de números aleatórios.

Complementando a seção, será apresentado o elenco de estatísticas a serem obtidas do universo de declarantes, quando da coleta da amostra.

No Capítulo 4 serão apresentados alguns resultados numéricos. A validação da amostra será objeto de discussão na Seção 4.1. As estatísticas do universo (totais, variâncias, expansores, etc.), apresentadas na Seção 4.2, servirão de subsídios para essa discussão. Os ganhos reais com a estratificação e com os especiais serão analisados a partir de medidas obtidas diretamente no universo.

Finalmente, no Capítulo 5, serão apresentadas algumas conclusões e indicações para possíveis extensões deste trabalho.

2 – HISTÓRICO

Ao longo dos últimos 12 anos o SERPRO vem utilizando técnicas de amostragem para obtenção de informações sobre alguns sistemas, como por exemplo: Imposto de Renda Pessoa Física – IRPF, Imposto de Renda Pessoa Jurídica – IRPJ, Imposto Territorial Rural – ITR, etc.

A experiência maior desenvolveu-se, entretanto, na área do IRPF, sendo que para este tributo o SERPRO dispõe de amostras estratificadas dos 12 últimos anos. Tais amostras foram usadas não só no Sistema Simulador de Legislações, como também em outros projetos que o SERPRO desenvolveu junto à SRF, como por exemplo: Malha de Suspeição do IRPF, Integração Parcial PF-PJ, etc.

A primeira amostra do IRPF foi dimensionada usando-se as Estatísticas Básicas do IRPF. Essas estatísticas fornecem, entre outras informações, total e freqüência, por classe de renda, da quase totalidade das variáveis envolvidas com o tributo.

Através de aproximações, calcularam-se, por classe de renda, as médias e as variâncias das variáveis envolvidas no dimensionamento da amostra.

Resultou deste processo uma amostra de aproximadamente 100 000 declarantes numa população de 6 000 000 (Exercício 1975). A partir de 1976 até 1978 usou-se sempre a amostra do ano $t-1$ para dimensionar a amostra no ano t.

Em 1979 desenvolveu-se um sistema que, além de rotinas que tratavam da escolha do número de estratos e dos limites ótimos para estratificação, oferecia ao usuário um módulo conversacional para alocação de amostras que funcionava de forma iterativa.

O sistema alocava a amostra e em seguida oferecia ao usuário a possibilidade de calcular os erros de estimação para algumas variáveis de interesse. Caso o erro amostral calculado para uma ou mais variáveis estivesse fora do aceitável, o sistema permitia ao usuário voltar ao início do processo e refazer a alocação.

Esse sistema foi desenvolvido em BASIC-PLUS para o PDP-11/40 e admitia como entrada médias e variâncias medidas em partições da população, definidas convenientemente pela variável de estratificação. No primeiro passo, o sistema escolhia o número de estratos e os limites ótimos desses estratos. Com a estratificação definida, o usuário entrava no módulo iterativo de alocação, até chegar a um resultado satisfatório em termos de tamanho de amostra e erros de estimativa das variáveis de interesse.

Com a experiência de dimensionamento, coleta e uso de amostras adquirida no período de 1975 até 1979 o SERPRO desenvolveu o pacote SANDA – Sistema de Análise de Dados.

O SANDA é um conjunto articulado de rotinas dotado de uma enorme facilidade para leitura de dados e para geração de tabelas e relatórios.

Além da parte relativa à amostragem, que inclui rotinas para determinação de estratos ótimos e alocação de amostras, o SANDA contempla a análise multivariada e a econometria (clássica e bayesiana).

A partir de 1979 todas as amostras do IRPF foram alocadas utilizando-se as técnicas de amostragem implementadas no SANDA.

3 – METODOLOGIA

3.1 – Estratificação

No processo de amostragem estratificada, a população-alvo é particionada em subpopulações menores, supostamente mais homogêneas com relação às características de interesse, denominadas estratos. Os estratos não possuem elementos comuns e sua reunião reproduz a população original.

Após a definição dos estratos, é feita uma seleção de amostras independentes em cada um deles, possivelmente segundo diferentes mecanismos probabilísticos.

Diversas razões podem levar ao uso de estratificação em amostragem, podendo-se mencionar :

- a conveniência administrativa, pela existência de subdivisões naturais da população;
- interesse em obterem-se estimativas com precisão predeterminada para subdivisões da população definidas previamente; e
- a possibilidade de obtenção de expressivos ganhos de precisão na estimação de características de interesse da população original.

No presente trabalho, o uso de estratificação deveu-se exclusivamente à última razão. Ganhos de precisão são obtidos através do emprego de amostragem estratificada quando a população é heterogênea, mas pode ser decomposta em subpopulações

homogêneas. A obtenção de estimativas em cada estrato, seguida da combinação para determinação da estimativa global, possibilita a eliminação da parcela mais significativa da variabilidade populacional, a existente entre os estratos. Esta questão ficará mais clara nas seções subseqüentes, quando se fizer menção às fórmulas dos estimadores.

3.1.1 – Variáveis de Estratificação

O uso tradicional de estratificação prevê a estimação de uma única característica de interesse e utiliza, na determinação dos estratos, a distribuição de freqüência da própria variável associada à característica, ou uma *proxy* correlacionada.

A amostra de declarantes do IRPF destina-se essencialmente a simulações de legislação tributária, sendo tipicamente uma amostra de múltiplos propósitos. Não existe, portanto, uma única característica de interesse, mas uma série delas, cuja importância nem mesmo pode ser determinada *a priori*.

A estratificação que objetive ganhos de precisão deve levar este aspecto em consideração, não podendo basear-se em uma única variável.

As variáveis que habitualmente são utilizadas nas simulações podem ser divididas em dois grupos: o das positivamente correlacionadas com a Renda Tributável do declarante (ex.: Renda Líquida, Imposto Líquido Devido, etc.), e o das pouco ou negativamente correlacionadas com a Renda Tributável (ex.: Renda Não-Tributável, Lucro Imobiliário, Participação Societária). Uma estratificação adequada às finalidades da amostra deve buscar a partição dos declarantes em grupos homogêneos segundo os dois conjuntos de variáveis.

Talvez a forma mais natural de se fazer isso seja através da estratificação cruzada da população com o uso de duas variáveis: uma que represente as variáveis do primeiro grupo e, outra, as do segundo grupo.

Com relação ao primeiro grupo, a variável mais adequada para uso na estratificação é a própria Renda Tributável - R. (Na determinação das especificações necessárias à seleção da amostra de declarantes para o exercício de um certo ano, faz-se sempre uso da amostra coletada no exercício anterior). Foram obtidas, ao todo, nove faixas de Renda Tributável para estratificação dos declarantes. Na Seção 3.1.1 é apresentada a metodologia utilizada para construção dos estratos de Renda Tributável. Os limites das diferentes faixas são ali apresentados.

Para o segundo grupo foi necessário introduzir-se uma variável artificial, obtida através da combinação de outras variáveis, consideradas de maior importância dentro do grupo.

São elas :

- Rendimento Não-Tributável (A_1)
- Imposto sobre Lucro na Alienação de Participação Societária (A_2)

- Imposto sobre Lucro na Alienação de Imóveis (A_3)
- Variação de Bens (A_4)
- Rendimento Tributável Exclusivamente na Fonte (A_5)

A cada uma das variáveis (A_1 a A_5) supracitadas associou-se um ponto de corte (C_1 a C_5), que deixa acima de si os valores mais expressivos da população (Na Seção 3.1.2 é apresentada a metodologia utilizada para determinação dos pontos de corte). A variável utilizada para estratificação dos declarantes, denominada Condição de Especial (E), foi definida a partir das variáveis A_1 a A_5 e dos respectivos pontos de corte C_1 a C_5. Sua expressão é dada a seguir:

$$E = \prod_{i=1}^{5} I_{[0,C_i]}(A_i)$$

onde

$$I_{[0,C_i]}(A_i) = \begin{cases} 1 & \text{se} \quad A_i < C_i \\ 0 & \text{se} \quad A_i \geq C_i \end{cases}$$

Observe-se que a variável E assume apenas dois valores: 1 (quando todas as variáveis são menores que os respectivos pontos de corte) e 0 (quando pelo menos uma das variáveis é maior que o seu ponto de corte).

Quando E for igual a 0, diremos que o declarante faz parte do grupo dos Especiais; quando E assumir o valor 1 ele será denominado Não-Especial.

O universo de declarantes foi, então, subdividido em 18 estratos, obtidos a partir do cruzamento entre as nove classes de Renda Tributável e os grupos de Especiais e Não-Especiais. Os critérios de seleção das amostras dentro dos estratos foram definidos de maneira diferenciada: indivíduos do grupo dos Especiais foram selecionados com probabilidade 1, independentemente da classe de Renda Tributável à qual pertencessem; nos demais estratos (i.é Não-Especiais através das nove classes de Renda Tributável) procedeu-se a uma Amostragem Binomial sem reposição (Na Seção 3.3 é apresentado o algoritmo de seleção da amostra nos estratos de Renda Tributável dos Não-Especiais).

3.1.2 – Definição dos Especiais

Na seção anterior, foi definida a variável E – Condição de Especial – utilizada, juntamente com a Renda Tributável, para estratificação do universo de declarantes do IRPF. A variável E divide a população em dois grupos: os Especiais e os Não-Especiais. Por definição, um declarante será considerado Especial quando apresentar valores expressivamente elevados em ao menos uma dentre as cinco variáveis de classificação citadas anteriormente. Nesta seção é descrita a metodologia utilizada para

obtenção dos pontos de corte das variáveis de classificação. Na especificação dos pontos de corte, duas questões fundamentais devem ser tratadas *a priori*:

- a definição do percentual de declarantes que irão compor o grupo dos Especiais, doravante representado por α; e
- o estabelecimento de pesos relativos para as variáveis de classificação na formação do grupo dos Especiais (i.é, deve-se definir, *a priori*, a importância de cada variável na classificação de um declarante como Especial).

O problema, então, pode ser colocado da seguinte maneira: desejamos encontrar valores C_1, C_2, C_3, C_4 e C_5 com as seguintes propriedades:

$$(i) \quad P\left(\bigcup_{i=1}^{5}(A_i \geq C_i)\right) = \alpha$$

$$(ii) \quad P(A_i \geq C_i) = P_i \cdot \delta, \qquad i = 1,2,\cdots,5$$

onde P_i representa o peso relativo da variável A_i na classificação dos Especiais e δ o percentual de indivíduos acima do ponto de corte de uma variável de classificação com peso unitário.

Observe-se que os pontos de corte são obtidos a partir de δ e das distribuições das variáveis de classificação. De fato, C_i é o percentil de ordem $1 - P_i \cdot \delta$ da distribuição de A_i.

O problema da obtenção dos pontos de corte C_i passa a ser:

i) a especificação das distribuições das variáveis de classificação; e

ii) a determinação do δ que satisfaça o sistema de equações anterior.

i) Distribuição das Variáveis de Classificação

Note-se, inicialmente, que todas as variáveis consideradas na classificação dos Especiais possuem em comum a característica de serem não-negativas, com forte assimetria na distribuição de freqüências (predominância de valores pequenos) e massa pontual no valor 0.

A hipótese assumida para as variáveis consideradas na classificação foi de que as restrições positivas das mesmas seguiam uma distribuição Log-Normal.

Em outras palavras, os logaritmos das restrições positivas das variáveis de classificação se distribuem segundo uma Normal.

Representamos a hipótese assumida por:

$$\log(A_i^*) \sim N(\mu_i, \sigma_i^2), \quad i = 1,2,\ldots,5$$

onde A_i^* representa a restrição de A_i a valores maiores que 0; μ_i e σ_i^2 são os parâmetros da distribuição (a esperança e a variância de $\log(A_i^*)$ respectivamente).

Sob a hipótese colocada acima, os pontos de corte se resumem à seguinte expressão:

$$C_i = \exp\{\mu_i + \sigma_i \cdot z(1 - (P_i \cdot \delta)/\gamma_i)\} \quad i = 1, 2, \ldots, 5$$

onde $\gamma_i = P(A_i > 0)$ e $z(P)$ representa o percentil de ordem P da distribuição $N(0,1)$.

Na prática, os parâmetros μ_i, σ_i^2 e γ_i são desconhecidos e devem ser substituídos na fórmula acima por estimativas.

Observe-se, assim, que os pontos de corte obtidos são, [illegible] estimativas dos verdadeiros pontos de corte.

Iremos representar estimadores de um determinado parâmetro θ, por $\hat{\theta}$ (assim $\hat{C}_i$ será o estimador de C_i etc.).

ii) Determinação de δ

Antes de tratar da determinação de δ propriamente dita, faremos duas observações importantes para a metodologia utilizada.

a) $\delta \in \left[\alpha / \sum_{i=1}^{5} P_i, \quad \alpha / P_m\right]$

onde $P_m = \max(P_1, P_2, P_3, P_4, P_5)$

Para ver isto note inicialmente que:

$$\begin{aligned} &\forall\, i = 1, 2, \ldots, 5 \quad P(A_i \geq C_i) = P_i \cdot \delta \leq \alpha \\ \Rightarrow\ &\forall\, i = 1, 2, \ldots, 5 \quad \delta \leq \alpha / P_i \\ \Rightarrow\ &\delta \leq \min(\alpha/P_1, \alpha/P_2, \alpha/P_3, \alpha/P_4, \alpha/P_5) = \alpha/P_m \end{aligned}$$

Por outro lado,

$$P\left(\bigcup_{i=1}^{5}(A_i \geq C_i)\right) = \alpha \leq \sum_{i=1}^{5} P(A_i \geq C_i) = \delta \cdot \sum_{i=1}^{5} P_i$$

$$\Rightarrow\ \delta \geq \alpha / \sum_{i=1}^{5}$$

b) A função $g(\delta) = P\left(\cup_{i=1}^{5}(A_i \geq C_i(\delta))\right)$ é contínua e crescente no intervalo $\left[\alpha / \sum_{i=1}^{5} P_i, \alpha/P_m\right]$ com $g\left(\alpha / \sum_{i=1}^{5} P_i\right) \leq \alpha$ e $g(\alpha/P_m) \geq \alpha$.

Para ver isto, note que:

$$\begin{aligned} g(\delta) &= 1 - P\left(\cap_{i=1}^{5}(A_i < C_i(\delta)\right) = \\ &= 1 - F(C_1(\delta), C_2(\delta), C_3(\delta), C_4(\delta), C_5(\delta)) \end{aligned}$$

onde F representa a função de distribuição conjunta do vetor $(A_1, A_2, A_3, A_4, A_5)$. Como já foi mencionado, $C_i(\delta)$ é na verdade o percentil de ordem $1 - P_i \cdot \delta$ da distribuição de A_i. Como tal, $C_i(\delta)$ é decrescente em δ, e como a F é crescente em cada um dos seus argumentos, segue-se que $g(\delta)$ cresce com δ.

Os valores da função nos extremos do intervalo são obtidos por substituição direta. A continuidade de $g(\delta)$ no intervalo decorre da hipótese de Log-Normalidade da distribuição das restrições positivas das variáveis de classificação.

As duas observações apresentadas acima implicam a existência de um valor, digamos δ^*, tal que:

$$g(\delta^*) = \alpha$$

ou seja, existe (pelo menos uma) solução para o problema.

Mais ainda, esta solução está necessariamente contida no intervalo

$$\left[\alpha / \sum_{i=1}^{5} P_i, \quad \alpha / P_m\right].$$

Especificada a distribuição das variáveis de classificação, com a estimação dos respectivos parâmetros associados (μ_i, σ_i^2 e γ_i), a obtenção de δ^* é feita iterativamente.

1) A partir de um valor inicial δ_n são obtidos os valores dos pontos críticos (estimados) usando:

$$\hat{C}_i = \exp\{\hat{\mu}_i + \hat{\sigma}_i \cdot z(1 - (P_i \cdot \delta_n)/\hat{\gamma}_i)\}$$

2) De posse destes valores, estima-se, por meio da amostra de declarantes disponível, o valor de $g(\delta_n)$.

3) Se $|g(\delta_n) - \alpha| \leq \xi$ o processo é interrompido e adota-se a solução:

$$\delta^* = \delta_n$$

onde ξ é a precisão desejada para a solução numérica.

Caso contrário, obtém-se um novo valor δ_{n+1}, a partir de δ_n, segundo algum procedimento numérico e volta-se para o passo 1.

Uma sugestão para o método de obtenção de δ_{n+1} a partir de δ_n é apresentada a seguir:

i) $a_o = \alpha / \sum_{i=1}^{5} P_i; \quad b_o = \alpha / P_m; \quad \delta_0 = (a_o + b_o)/2$

ii) se $g(\delta_n) > \alpha$ então :

$\delta_{n+1} = \delta_n - [(g(\delta_n) - \alpha)/(g(\delta_n) - g(a_n))] \cdot (\delta_n - a_n)$

$a_{n+1} = a_n; \quad b_{n+1} = \delta_n$

se $g(\delta_n) < \alpha$ então

$$\delta_{n+1} = \delta_n + [(\alpha - g(\delta_n))/(g(b_n) - g(\delta_n))] \cdot (b_n - \delta_n)$$
$$a_{n+1} = \delta_n \quad b_{n+1} = b_n$$

Os pontos de corte adotados para definição dos Especiais serão:

$$\hat{C}_i = \exp\{\hat{\mu}_i + \hat{\sigma}_i \cdot z(1 - P_i \cdot \delta^*/\hat{\gamma}_i)\} \quad i = 1, 2, \cdots, 5$$

Observe-se que os valores obtidos, como se baseiam em uma amostra de declarantes do exercício imediatamente anterior àquele para o qual a nova amostra será selecionada, necessitam ser corrigidos para compensar a inflação do período.

Assim, os pontos de corte a serem efetivamente adotados na classificação dos Especiais serão dados por:

$$\hat{C}_i^c(\delta^*) = \hat{C}_i(\delta^*) \times I$$

onde I é um índice de variação de preços do exercício atual em relação ao passado (IPC anual p.e.).

A metodologia apresentada nesta seção foi, com algumas adaptações, aplicada na obtenção dos pontos de corte para estratificação dos declarantes do exercício fiscal de 1986.

Nesta aplicação adotou-se:

$$\alpha = 16.000/7.000.000 = 2.286 \times 10^{-3}$$
$$P_1 = P_5 = 2$$
$$P_2 = P_4 = 1.1$$
$$P_3 = 1$$

Os parâmetros μ_i, σ_i e γ_i foram estimados por método de momentos a partir da amostra de declarantes do exercício de 85.

Pelas observações feitas anteriormente, temos que o valor δ necessariamente está contido no intervalo:

$$[0.0023/(2 + 2 + 1 \cdot 1 + 1 \cdot 1 + 1), 0.002286/2]$$
$$= [3.194, 11.43] \times 10^{-4}$$

De fato, obteve-se ao final dos procedimentos numéricos que:

$$\delta^* = 4.229 \times 10^{-4}$$

Os valores encontrados para os pontos de corte, já com valores corrigidos para a inflação do período, utilizou-se para correção o INPC referente ao ano de 1985, são apresentados na Tabela 1.

Na Tabela 1, n_i representa o número de indivíduos na população com valores de A_i acima do ponto de corte $\hat{C}_i^c$. Note-se que a soma dos n_i não é igual ao número de Especiais.

Tabela 1
Pontos de Corte para Definição dos Especiais segundo as Variáveis de Classificação

i	$\hat{C}_i^c(Cr\$\,85)$	n_i	$\% \times 10^{-4}$
1	2 035 442 000	5 940	8 486
2	964 343	3 390	4 843
3	18 227 880	2 960	4 229
4	3 164 460 000	3 420	4 886
5	386 301 500	5 840	8 343
Total		15 982	22 830

3.1.3 – Construção dos Estratos

Conforme já abordado anteriormente no texto, a população de declarantes foi estratificada segundo duas variáveis: a Renda Tributável e a Condição de Especial. Na seção precedente foi exposta a metodologia utilizada para construção dos estratos de declarantes Especiais e Não-Especiais, com base na segunda variável citada. Nesta seção é apresentada a metodologia de estratificação segundo a Renda Tributável, adotada neste estudo.

A construção de estratos em uma população envolve, de um modo geral, três questões básicas: a escolha da variável de estratificação, a determinação dos limites e da quantidade de estratos.

i) Variável de Estratificação

Idealmente a população deveria ser estratificada segundo a Renda Tributável do exercício para o qual a amostra está sendo colhida. Como esta informação só fica disponível no momento da seleção da amostra, torna-se necessário utilizar outra variável, correlacionada com a ideal, que esteja a disposição para os cálculos necessários. Em face da grande estabilidade da distribuição de Renda Tributável no Brasil, a escolha recaiu sobre a Renda Tributável do exercício anterior ao da seleção da amostra. Os limites entre os estratos assim obtidos necessitaram sofrer correção para compensar a inflação entre os exercícios. Utilizou-se para tal o INPC acumulado no ano.

ii) Limites dos Estratos

A metodologia utilizada no estudo para determinação de limites ótimos entre estratos é apresentada em Cochran (1977, pág. 129-130), seguindo trabalho de Dalenius e Hodges (1959). Esta metodologia, que denominaremos Método das Freqüências, foi implementada no SANDA, fazendo parte de sua sub-rotina de estratificação. O Método das Freqüências fornece uma aproximação para a solução do problema de determinação dos limites ótimos entre estratos, com o atrativo de ser computacionalmente simples.

Resumidamente podemos descrever o método da seguinte forma: uma população de N indivíduos deve ser estratificada em L classes, segundo a variável y.

Sejam:

N_h = número de indivíduos do estrato h;

$N = \sum_{h=1}^{L} N_h$, número de indivíduos na população;

$W_h = N_h/N$, o peso relativo do estrato h;

$\bar{Y}_h = \sum_{i=1}^{N_h} y_{hi}/N_h$, a média da característica y no estrato h (y_{hi} refere-se ao valor da característica y para o i-ésimo indivíduo do estrato h);

$S_h^2 = \sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_h)^2/(N_h - 1)$, a variância do estrato h;

n_h = tamanho da amostra no estrato h;

$\bar{y}_h = \sum_{i=1}^{n_h} y_{hi}/n_h$, a média amostral da característica y no estrato h;

$\bar{y}_{st} = \sum_{h=1}^{L} W_h \cdot \bar{y}_h$, o estimador da média da característica y na população, baseado na amostra estratificada.

Então, sob alocação de Neyman (i.é $n_h = n \cdot (W_h \cdot S_h)/(\Sigma_h W_h \cdot S_h)$) temos:

$$V(\bar{y}_{st}) = (1/n) \cdot \left(\sum_{h=1}^{L} W_h \cdot S_h \right)^2 - (1/N) \cdot \sum_{h=1}^{L} W_h \cdot S_h^2$$

Se a população for suficientemente grande para ignorar-se o fator de correção para populações finitas temos:

$$V(\bar{y}_{st}) = (1/n) \cdot \left(\sum_{h=1}^{L} W_h \cdot S_h \right)^2$$

O problema da determinação dos limites ótimos dos estratos consiste então na obtenção de valores $y_1, y_2, \cdots, y_{L-1}$ tais que

$$\sum_{h=1}^{L} W_h \cdot S_h$$

seja mínima.

A metodologia proposta por Dalenius e Hodges baseia-se na hipótese de que os estratos são numerosos e estreitos, de tal forma que a densidade de y, $f(y)$, é aproximadamente constante em cada estrato. Neste caso, segue-se que:

$$W_h = \int_{y_{h-1}}^{y_h} f(t)dt \cong f_h \cdot (y_h - y_{h-1})$$

e

$$S_h \cong (1/\sqrt{12}) \cdot (y_h - y_{h-1})$$

onde f_h representa o valor (constante) de $f(y)$ no estrato h.

Decorre então que:

$$\sqrt{12} \cdot \sum_{h=1}^{L} W_h \cdot S_h \cong \sum_{h=1}^{L} f_h \cdot (y_h - y_{h-1})^2 = \sum_{h=1}^{L} \left[\sqrt{f_h} \cdot (y_h - y_{h-1})\right]^2$$

Fazendo

$$Z(y) = \int_{y_0}^{y} \sqrt{f(t)}dt$$

então

$$Z(y_h) - Z(y_{h-1}) = \int_{y_h-1}^{y_h} \sqrt{f(t)}dt \cong \sqrt{f_h} \cdot (y_h - y_{h-1})$$

Logo,

$$\sqrt{12} \cdot \sum_{h=1}^{L} W_h \cdot S_h \cong \sum_{h=1}^{L} (Z_h - Z_{h-1})^2$$

Como $(Z_L - Z_0)$ é fixo, segue-se que a soma acima é minimizada fazendo-se $(Z_h - Z_{h-1})$ constante. A regra para obtenção dos limites dos estratos é, então, buscar valores tais que a acumulada de $\sqrt{f(y)}$ seja dividida em L valores iguais.

iii) Número de Estratos

A determinação da quantidade ótima de estratos envolve, habitualmente, considerações sobre a precisão mínima desejada e o custo adicional acarretado pela inclusão de novos estratos na partição da população.

No presente trabalho, o custo não é um entrave para a estratificação, já que toda a população de declarantes está disponível em meio magnético. O problema maior, que contra-indica a subdivisão excessiva da população, associa-se ao uso da Renda Tributável do exercício anterior como variável de estratificação, seguida de correção dos limites segundo um índice de preços. Este procedimento pode acarretar distorções na classificação dos declarantes do exercício atual, sendo o problema tão mais crítico quanto maior a quantidade de estratos.

Tabela 2
Erro Relativo do total Estimado
para a Variável de Estratificação

Número de Estratos

Erro Relativo	01	02	03	04	05	06	07	08	09	10
$\times 10^{-1}$	10.51	5.76	4.13	3.33	2.65	2.21	1.98	1.72	1.54	1.42

Erro Relativo	11	12	13	14	15	16	17	18	19	20
$\times 10^{-1}$	1.34	1.22	1.15	1.09	1.03	1.01	0.95	0.92	0.88	0.80

A metodologia empregada para especificação do número de estratos adequado foi a seguinte:

- foram obtidos os limites ótimos, segundo o Método das Freqüências exposto no item anterior, para uma grande variedade de número de estratos (1 a 20);
- para cada partição ótima, associada a um determinado número de estratos, foi estimado o erro relativo do estimador da variável de estratificação; e
- analisaram-se os decréscimos nos erros relativos obtidos com o aumento do número de estratos, determinando-se o valor a partir do qual os ganhos marginais de precisão eram desprezíveis. Este foi o número utilizado na construção dos estratos.

A Tabela 2 apresenta a variação dos erros relativos das estimativas de total da variável de estratificação para as diversas quantidades de estratos considerados.

Analisando-se a Tabela 2, verifica-se que a partir de nove estratos há pequena diminuição dos erros relativos, tendo-se decidido por este número na estratificação final da população.

Os limites dos estratos, obtidos pelo Método das Freqüências, já corrigidos para cruzeiros de 1985, são apresentados na Tabela 3.

3.2 – Alocação

3.2.1 – Metodologia

A alocação da amostra aos estratos é feita prevendo-se critérios de seleção diferenciados para os estratos de Renda Tributável e de Condição de Especial. Os indivíduos

Tabela 3
Limites de Estratificação da Variável
Renda Tributável para 1985

Estrato	Limites (Cr$ 1985)
01	0 – 12983750
02	12 983 751 – 22 106 500
03	22 106 501 – 28 866 500
04	28 866 501 – 38 025 000
05	38 025 001 – 49 302 500
06	49 302 501 – 74 620 000
07	74 620 001 – 109 102 500
08	109 102 501 – 165 360 000
09	+ de 165 360 000

que compõem o estrato dos Especiais são selecionados com probabilidade 1. O restante da amostra é distribuído entre os Não-Especiais através dos estratos de Renda Tributável , segundo alocação de Neyman.

A alocação de Neyman objetiva minimizar a variância do estimador da média, ou do total populacional, de uma dada característica. Assumindo-se que a população está dividida em L estratos e seguindo-se a mesma notação mencionada no item ii) de 3.1.3, então, para um determinado tamanho total de amostra n, tem-se a alocação de Neyman dada por:

$$n_h = \left(W_h \cdot S_h \Big/ \sum_{h=1}^{L} W_h \cdot S_h \right) \cdot n = \left((N_h \cdot S_h) \Big/ \sum_{h=1}^{L} N_h \cdot S_h) \right) \cdot n$$

O aspecto mais importante a ser observado é que o tamanho relativo da amostra no estrato aumenta com o seu peso e a variabilidade da característica sob estudo.

A amostra de declarantes do IRPF não se destina a estimar uma única característica populacional, mas a um conjunto delas. Desta forma é desaconselhável a escolha de uma única variável para fins de alocação, já que isso poderia implicar a má estimação de outras variáveis pouco correlacionadas com ela. A solução adotada foi a de efetuar um conjunto de alocações separadas para um grupo de variáveis-alvo (Foram utilizadas, além das variáveis já mencionadas na seção de estratificação, as seguintes variáveis de interesse: Rendimento Cédula - G, Abatimento Médico, Renda Líquida, Imposto

Tabela 4
Resultados da Alocação da Amostra segundo
Faixas de Renda Tributável

Estrato	N_h	n_h	$f_h(\times 10^{-2})$	Nº Esperado de Especiais
01	739 800	3 219	0,4351	378
02	1 878 000	5 406	0,2879	319
03	1 317 000	2 829	0,2148	2 269
04	1 070 000	3 182	0,2974	1 010
05	708 100	2 574	0,3635	533
06	724 300	5 851	0,8078	2 001
07	396 500	4 290	1,0820	2 287
08	184 900	3 221	1,7420	2 590
09	75 650	14 430	19,0746	4 595
Total	7 094 250	45 000		15 982

Líquido Devido e Redução Investimento), consideradas mais importantes. A alocação final é obtida a partir da combinação destas alocações segundo o esquema abaixo.

Sejam:

$n_h(p)$ - o tamanho da amostra no estrato h, com base na alocação de Neyman para a variável p;

n'_h - o maior tamanho de amostra para o estrato $h (= \max_P(n_h(p)))$;

então, a alocação final é dada por:

$$n_h = \left[n'_h \Big/ \left(\sum_{h=1}^{L} n'_h \right) \right] \cdot n$$

Todas estimativas necessárias aos cálculos descritos anteriormente foram efetuadas com base na amostra do exercício anterior, utilizando-se o SANDA.

3.2.2 – Resultados da Alocação

Com base na experiência anterior, definiu-se que uma amostra de 61 000 declarantes seria suficiente para fornecer resultados com precisão aceitável. A parcela da amostra

Tabela 5
Ganhos de Precisão para as Variáveis de Interesse

Variável	$Var(a)$ $\times 10^2 2$	$Var(b)$ $\times 10^2 2$	$Var(c)$ $\times 10^2 2$	G_1	G_2
RTRIB	1.647	1.569	32,21	1,05	48,70
RLIQUIDA	1.104	925,90	70,34	1,19	13,20
ILDEV	254	151,70	4,38	1,67	34,60
RNTRIB	92.510	7.805	7.531	11,85	1,04
VARBENS	95.100	7.861	6.791	12,10	1,16
LIMOVEL	1,91	0,20	0,33	9,37	0,60
PSOCIE	5,87	$6,08 \times 10^{-5}$	$5,96 \times 10^{-5}$	9.657	1,02
RTXFONT	3.124	200,20	199,60	15,60	1,00
RCEDG	35,35	14,37	10,23	2,46	1,41
ABATMED	3,15	3,97	2,54	0,80	1,56
REDINV	0,025	0,03	0,0081	0,83	3,74

referente ao estrato dos Especiais possui tamanho aleatório, já que definidos os critérios de classificação dos indivíduos em Especiais e Não-Especiais, o tamanho exato só é conhecido após a seleção. Determinou-se, também com base na experiência pregressa, que um tamanho esperado de 16 000 Especiais seria suficiente.

O restante da amostra (45 000 declarantes) foi alocado aos estratos segundo a metodologia descrita no item anterior, tendo-se obtido os seguintes resultados constantes da Tabela 4.

3.2.3 – Validação Preliminar

A partir da amostra do exercício anterior, foi feita uma validação preliminar da amostra obtida. As variâncias dos estimadores dos totais de cada uma das variáveis-alvo, e algumas outras também de interesse, foram estimadas, com base na alocação obtida, segundo três situações:

a) sem estratificação nenhuma;
b) estratificando-se apenas pela Condição de Especial; e
c) com a estratificação completa.

Tabela 6
Coeficientes de Variação para as Variáveis
de Interesse, supondo Estratificação Completa

Variável	Coef. de Variação ($\times 10^{-2}$)
RENDA TRIBUTÁVEL	0,0699
RENDA LíQUIDA	0,1690
IMPOSTO LÍQ. DEVIDO	0,2930
RENDA NÃO-TRIBUTÁVEL	1,3700
VARIAÇÃO DE BENS	1,1600
LUCRO IMOBILIÁRIO	6,2800
PARTICIPAÇÃO SOCIETÁRIA	0,3060
REND. TRIB. EXCLUS. NA FONTE	0,2460
REND. CÉDULA-G	0,3890
ABAT. MÉDICO	1,3100
REDUÇÃO INVESTIMENTO	0,6893

Os ganhos associados à estratificação completa e ao uso do estrato dos Especiais foram estimados a partir das razões entre as variâncias das situações a e b, e b e c, respectivamente. Os valores obtidos são listados a seguir.

Sejam:

$Var(a)$ - Variância da média na situação a

$Var(b)$ - Variância da média na situação b

$Var(c)$ - Variância da média na situação c

$G_1 = Var(a)/Var(b)$

$G_2 = Var(b)/Var(c)$

Observe-se que as variáveis positivamente correlacionadas com a Renda Tributável apresentam ganhos consideráveis devido à estratificação por aquela variável, tendo pouca influência o uso do estrato de Especiais. Com as variáveis pouco correlacionadas com ela ocorre o oposto, sendo expressivo apenas o ganho associado à inclusão do estrato dos Especiais.

Todas as variáveis analisadas apresentaram coeficientes de variação bastante reduzidos, tendo-se obtido as seguintes estimativas constantes da Tabela 6.

A validação da amostra com dados reais, ou seja, para o exercício para o qual ela foi planejada, é apresentada no próximo capítulo.

3.3 – Seleção da Amostra

3.3.1 – Programa Amostrador

A seleção da amostra é realizada pelo Programa Amostrador, que, além da seleção propriamente dita, gera estatísticas do universo necessárias para a validação da amostra e determinação das variâncias dos estimadores.

Ao Programa Amostrador são fornecidas as seguintes informações:

- Limites dos estratos de Renda Tributável;
- Definição dos estratos dos Especiais;
- Frações amostrais (F_A) (por estrato); e
- Definição das estatísticas do universo.

De posse dessas informações e usando uma rotina de geração de números aleatórios previamente testada, o programa executa o seguinte procedimento operacional:

1 - Lê um registro associado a um declarante no universo de declarantes do IRPF;

2 - A partir das informações do declarante determina o estrato h ao qual ele pertence;

3 - Sorteia um número aleatório U, com distribuição uniforme no intervalo [0,1];

4 - Se $U \leq F_A(h)$ seleciona o declarante para a amostra (grava registro no arquivo amostra);

5 - Processa rotina de estatísticas do universo; e

6 - Repete o procedimento até o final do arquivo de entrada (universo de declarantes do IRPF).

3.3.2 – Procedimento Amostral

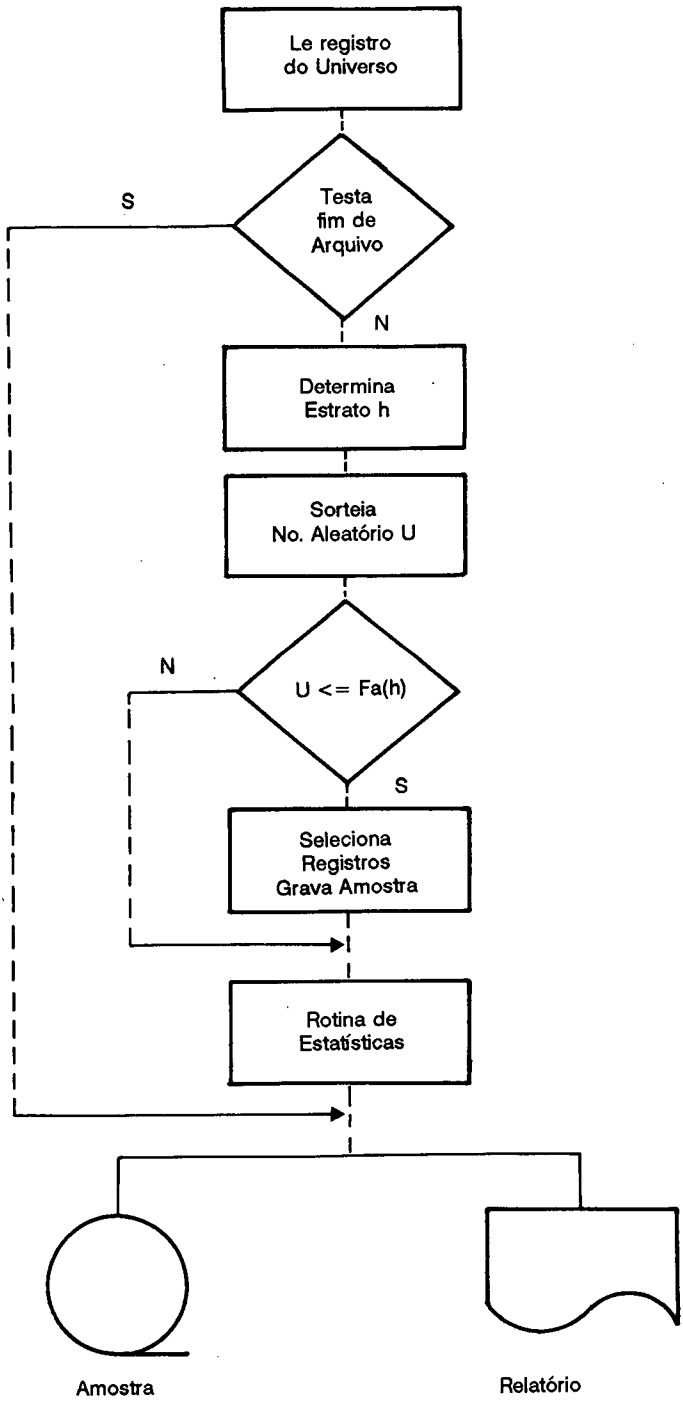
O procedimento amostral efetua a seleção segundo o esquema Binomial, levando a uma amostra sem reposição da população de declarantes. O universo (Cadastro de Declarantes do IRPF) é um arquivo em fita magnética que é lido seqüencialmente pelo Programa Amostrador. Para cada registro lido (associado a um declarante) o programa executa o procedimento operacional apresentado anteriormente, até que o

final do arquivo é encontrado. A rotina de geração de números aleatórios (RANDOM) gera valores segundo uma distribuição Uniforme no intervalo [0,1].

Para se obter uma fração amostral esperada α, em um determinado estrato h, basta que comparemos o número aleatório U sorteado no intervalo [0,1] com a fração amostral esperada α daquele estrato.

Ou seja, se para $U \leq F_A(h)$ o indivíduo do estrato h é selecionado, então o estrato h terá uma fração amostral aproximadamente igual a $F_A(h)$. Procedendo-se desta forma, não garantimos que seja selecionada uma amostra de tamanho n. O que se obtém na prática é uma amostra de tamanho $n' \cong n$, com frações amostrais $F'_A(h)$, próximas das frações amostrais teóricas $F_A(h)$ desejadas. As frações amostrais F'_A são variáveis aleatórias cuja esperança é igual à fração teórica. Como o número de declarantes por estrato é bastante elevado, espera-se que a fração $F'_A(h)$ encontre-se bem próxima da fração teórica $F_A(h)$), $h = 1, 2, \cdots, L$.

Fluxograma Simplificado



O procedimento é reexecutado até que o final do Arquivo de Entrada é encontrado.

Tabela 7
Frações Amostrais Esperadas e Observadas segundo Estratos de Renda Tributável

Estrato	Fração Amostral Teórica(F_A)	Fração Amostral Real(F'_A)	Erro Relativo $\frac{(F_A - F'_A)}{F_A}$
01	0,004351	0,004327	0,0055
02	0,002879	0,002848	0,0108
03	0,002148	0,002194	−0,0209
04	0,002974	0,002924	0,0170
05	0,003635	0,003627	0,0022
06	0,008078	0,008033	0,0056
07	0,010820	0,010770	0,0046
08	0,017420	0,017348	0,0041
09	0,190746	0,188806	0,0102
Especiais	1,000000	1,000000	0,0000

erro máximo = 2,09%; erro mínimo = 0,22%

As diferenças entre as frações amostrais teóricas e reais são exibidas na Tabela 7.

Poderiam ter sido adotados procedimentos de seleção que garantissem um tamanho de amostra n especificado na alocação. Tais procedimentos normalmente envolvem um consumo excessivo de tempo de processamento (CPU-TIME), o que no caso em questão levaria a um custo muito elevado para a seleção da amostra, dado o tamanho da população (8 330 000 declarantes).

Dado o exposto, optou-se pelo procedimento de seleção que procura manter as frações amostrais teóricas especificadas na alocação. Observando a Tabela 7 percebe-se que os erros relativos entre as frações amostrais teórica e real por estrato são aceitáveis, pois variam (em módulo) entre 0,22% e 2,09% com um valor médio ($\sim$ 0,8%) inferior a 1%.

Como já foi mencionado anteriormente, o dimensionamento da amostra do ano t é feito usando-se a amostra do ano $t-1$. Assim, o dimensionamento da amostra-86 foi realizado usando-se a amostra-85. Os valores monetários são convenientemente corrigidos e, levando-se em conta o fato da população de declarantes ser muito estável (pouca migração entre os estratos de Renda Tributável, o perfil de renda dos contri-

Tabela 8
Tamanho da Amostra-85 segundo os Estratos

Estrato	N_h	n_h	Especiais
01	739 800	3 219	378
02	1 878 000	5 406	319
03	1 317 000	2 829	2 269
04	1 070 000	3 182	1 010
05	708 100	2 574	533
06	724 300	5 851	2 001
07	396 500	4 290	2 287
08	184 900	3 221	2 590
09	75 650	14 430	4 595
Total	7 094 250	45 000	15 982
População	7 110 232	Amostra	60 982

buintes não se altera substancialmente em um ano, etc.) espera-se que o desenho da amostra que é "ótimo" em $t-1$ (85) esteja próximo do "ótimo" em t (86).

Dado que há um crescimento natural (crescimento vegetativo) da população de declarantes, e como o procedimento de seleção é tal que procura-se manter não o tamanho n total da amostra, mas as frações amostrais teóricas $F_A(h), h = 1,2,\cdots,L$ resultantes do dimensionamento, o que se obtém na seleção é uma amostra de tamanho $n' > n$.

No caso específico do dimensionamento da amostra-86, usou-se uma amostra de painel de declarantes do IRPF entre os anos 83 e 85.

Assim, a partir de 83 a amostra acompanhou os mesmos declarantes ao longo de três anos consecutivos (83,84 e 85).

Tal procedimento permitiu estudos mais detalhados (e precisos) sobre os impactos produzidos por alterações na legislação do IRPF, principalmente no que se refere à evolução da progressividade do tributo no período de 83 a 85.

Isto posto, apresenta-se a seguir,para efeito de comparação, duas tabelas. A primeira (Tabela 8) contém o resultado do dimensionamento obtido a partir da amostra-85 (ver Seção 3.2).

Tabela 9
Tamanho da Amostra-86 segundo os Estratos

Estrato	N_h	n_h	Especiais
01	1 034 408	4 476	1 658
02	2 431 930	6 927	1 216
03	1 341 690	2 944	958
04	1 099 504	3 215	1 248
05	769 701	2 792	1 408
06	838 741	6 738	2 757
07	447 153	4 816	3 095
08	245 494	4 259	3 301
09	107 041	20 210	6 958
Total	8 315 702	56 377	22 599
População	8 338 261	Amostra	78 976

O tamanho total da amostra dimensionada (60 982) resulta em uma fração amostral global de 0,00857 (0,857%) .

Na Tabela 9 estão as medidas obtidas na seleção da amostra-86.

Comparando-se os totais da população de declarantes em 83 (7 110 232) e 86 (8 338 261) percebe-se que houve um crescimento de 17,27% neste período.

Como as frações amostrais foram mantidas (ver Tabela 7) pelo procedimento de seleção adotado, obteve-se uma amostra de tamanho $n' = 78.976$. A fração amostral global ficou em 0,00947 (0,947%), ligeiramente superior à teórica (0,857%).

3.3.3 – Estatísticas do Universo

A rotina de estatísticas do universo calcula basicamente o total e a soma de quadrados por estrato das seguintes variáveis de interesse:

- Rendimento Tributável;
- Rendimento Cédula-G;
- Renda Líquida;
- Abatimento Médico;

- Redução Investimento;
- Imposto Líquido Devido;
- Imposto s/ Lucro na Alienação de Participações Societárias;
- Imposto s/ Lucro na Alienação de Imóveis;
- Rendimentos Não-Tributáveis;
- Rendimentos Tributados Exclusivamente na Fonte; e
- Variação Patrimonial.

Os totais serão usados na fase de validação da amostra, onde serão comparados os valores obtidos no universo com os valores estimados a partir da amostra.

A soma de quadrados será útil para calcularmos a variância dos estimadores dos totais das características citadas e para verificação dos ganhos obtidos com a estratificação e com os Especiais.

Além disso, calculou-se para cada estrato h o tamanho da população do estrato no universo (N_h) e na amostra (n_h). Esses números serão utilizados para fazer a expansão da amostra.

No capítulo seguinte serão apresentados alguns resultados numéricos, acompanhados de uma análise sobre a validação da amostra . Serão vistos ainda os ganhos relativos à estratificação e à presença dos Especiais.

4 – VALIDAÇÃO DA AMOSTRA

Após a etapa de seleção da amostra de declarantes do IRPF, alguns procedimentos de validação são utilizados, com o intuito de auferir a qualidade das informações coletadas. Tais procedimentos são de dois tipos: cálculo dos erros relativos de estimação (a partir da amostra selecionada) dos totais e coeficientes de variação dos estimadores destes totais, para um conjunto de variáveis-alvo; e a determinação dos ganhos (em termos de redução das variâncias dos estimadores dos totais das variáveis-alvo) obtidos por meio dos procedimentos de estratificação adotados.

4.1 – Erros Relativos

Os erros relativos de estimação são calculados como a diferença percentual entre os valores estimado e real para o parâmetro considerado, com relação ao valor real. Ou seja, se θ é o parâmetro de interesse e $\hat{\theta}$ o seu valor estimado a partir da amostra, o erro relativo de estimação é definido como se segue:

$$e(\hat{\theta}) = ((\hat{\theta} - \theta) \Big/ \theta) \cdot 100$$

Tabela 10
Erros Relativos do Total por Estrato

RENDG	RTRIB	AMED	RLIQ	REDINV	ILD	PSOCIE	LIMOV	RNTRIB	RTXF	VBENS
11.57	0.05	3.03	2.61	−28.76	20.49	*	36.79	−2.99	14.50	−5.27
−11.92	0.06	−4.11	0.01	8.55	−1.47	418.47	90.12	−7.03	−17.25	−1.62
−3.77	−0.07	2.80	−0.95	−5.10	−2.83	52.54	−82.63	−0.74	−6.00	0.87
−18.64	−0.07	−1.01	−0.21	2.63	−0.50	*	118.01	−5.76	13.95	0.06
−4.49	−0.12	−2.93	−0.02	−0.23	0.02	*	154.23	−3.41	5.64	−5.52
−3.11	−0.28	−1.63	−0.02	1.80	−0.12	8.37	−15.76	0.93	2.78	−0.89
6.85	0.08	−0.96	0.23	0.35	0.35	−27.97	−2.25	−1.22	−8.19	−1.16
6.87	−0.18	1.44	−0.18	1.57	−0.24	−51.01	24.67	0.91	−2.66	0.91
−0.32	−0.02	2.45	−0.01	0.26	−0.11	−16.79	10.36	1.20	−0.60	1.24
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
−2.39	−0.08	−0.39	−0.05	0.89	−0.11	−0.01	7.93	−1.17	−0.21	−0.82

Nota - Os erros relativos por estrato, associados à Participação Societária, não têm maior significado, já que pela regra de definição dos Especiais a quase totalidade (98,85%) dos indivíduos com valor positivo nesta variável são considerados Especiais.

No trabalho os parâmetros de interesse são o total e o coeficiente de variação do estimador do total para um grupo de variáveis-alvo. As estimativas foram obtidas a partir da amostra e os valores reais foram determinados durante a etapa de seleção da amostra, quando todo o universo de declarantes foi varrido pelo Programa Amostrador. Os erros relativos, por estrato e global, para os parâmetros e variáveis de interesse são apresentados nas Tabelas 10 e 11.

Observe-se que, de um modo geral, os erros relativos são mais expressivos nas faixas de baixa Renda Tributável, sendo que os totais são estimados de forma mais precisa que os coeficientes de variação. O próprio procedimento de estratificação e alocação explica os resultados obtidos, já que foi dada prioridade aos estratos de maior variabilidade (faixas de Renda Tributável mais altas) e visou-se à minimização da variância dos estimadores dos totais. Note-se também que, em geral, os erros relativos são menores para as variáveis utilizadas na estratificação e na alocação da amostra.

Analisando-se apenas a linha referente aos erros relativos dos totais para a população como um todo, conclui-se que a amostra é de ótima qualidade para os

Tabela 11
Erros Relativos do C.V.Estimado por Estrato

RENDG	RTRIB	AMED	RLIQ	REDINV	ILD	PSOCIE	LIMOV	RNTRIB	RTXF	VBENS
−5.62	0.72	−39.17	−58.39	−15.35	6.20	*	−26.28	−14.23	2.46	−17.34
−5.69	2.05	−48.85	−13.28	−8.46	−1.77	−61.13	−21.65	−7.23	2.77	−11.15
2.27	3.60	−22.26	−13.63	1.78	1.12	−61.45	−40.60	0.77	35.50	13.71
9.34	0.84	−30.26	−2.39	−2.49	−1.51	*	−30.66	−5.80	1.53	9.01
0.23	6.10	−10.02	−4.24	0.18	0.64	*	−37.88	−4.41	1.27	−10.76
0.79	5.98	−18.05	−3.52	−1.09	0.40	5.36	−4.94	5.42	−2.64	2.29
−0.42	4.64	−5.51	−0.42	−1.60	−0.18	−4.57	5.76	−2.86	6.70	−3.32
−2.18	4.61	9.88	−6.78	−1.86	1.48	−8.02	−9.56	−0.74	2.68	2.58
−0.81	−1.82	70.38	−1.01	−1.24	−0.86	11.18	−2.01	0.07	−1.32	−0.34
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
−0.56	1.70	−19.91	−14.97	−1.20	−0.11	−1.82	12.30	−4.81	9.05	−0.83

Nota - Os erros relativos para os Especiais são nulos, já que todos os indivíduos deste grupo são selecionados na amostra.

propósitos de simulação a que se destina.

4.2 – Ganhos Associados à Estratificação

Conforme mencionado na introdução da seção, os ganhos são obtidos a partir da comparação entre as variâncias dos estimadores dos totais das variáveis-alvo, com e sem o uso da estratificação.

Como a população de declarantes foi estratificada segundo o cruzamento das estratificações por duas variáveis, torna-se interessante a análise dos diferentes ganhos associados às diversas possibilidades de estratificação.

Com este objetivo, são introduzidas abaixo as seguintes variâncias dos estimadores:

V_1 - Variância com amostra aleatória simples sem reposição;

V_2 - Variância com estratificação pelos Especiais;

V_3 - Variância com a estratificação por Renda Tributável; e

V_4 - Variância com estratificação cruzada.

Tabela 12
Realocação da Amostra sem os Especiais

Estrato	Antiga	Nova
01	4476	7093
02	6927	6795
03	2944	5089
04	3215	5234
05	2792	4251
06	6738	10296
07	4816	9091
08	4259	5570
09	20210	25657
Especiais	22599	0
Total	78976	78976

No cálculo das variâncias acima, utilizou-se sempre o mesmo tamanho de amostra, tendo sido necessário, no caso de V_3, refazer a alocação para distribuir, entre os estratos de Renda Tributável, a parcela da amostra representada pelos Especiais. As duas alocações são apresentadas na Tabela 12.

Os ganhos considerados na análise são descritos a seguir:

$G_1 = V_1/V_2$ (Ganho com estratificação por Condição de Especial);
$G_2 = V_1/V_3$ (Ganho com estratificação por Rendimento Tributável);
$G_3 = V_1/V_4$ (Ganho com estratificação cruzada);
$G_4 = V_2/V_4$ (Ganho adicional da estratificação por Rendimento Tributável); e
$G_5 = V_3/V_4$ (Ganho adicional da estratificação por Condição Especial).

A Tabela 13 apresenta os ganhos obtidos para as variáveis-alvo.

O ganho G_3 é o mais relevante para julgamento da qualidade da amostra. Por ele pode-se concluir pela excelência da amostra, já que a quase totalidade dos ganhos foram bastante expressivos (a única exceção é a variável Abatimento Médico).

O ganho associado à Participação Societária é explicado pelo fato do critério para classificação dos Especiais levar praticamente a um censo dos indivíduos que possuem

Tabela 13
Ganhos Associados à Estratificação

Var. / Ganhos	G_1	G_2	G_3	G_4	G_5
REND. CÉDULA -G	2.05	2.83	2.18	1.06	0.77
REND. TRIB.	0.58	42.14	48.21	83.27	1.14
ABAT. MÉDICO	0.84	1.30	1.12	1.33	0.86
RENDA LÍQUIDA	0.74	13.66	11.68	15.79	0.85
REDUÇÃO INVEST.	0.72	4.45	2.86	3.99	0.64
IMP. LÍQ. DEVIDO	1.38	33.94	51.25	37.01	1.51
IMP.LUCRO PSOCIE	40.5×10^4	8.56	35.8×10^4	0.88	41.8×10^3
IMP.LUCRO IMOB.	31.50	1.26	28.70	0.91	22.84
REND.NÃO TRIB.	40.77	3.12	45.44	1.11	14.59
REND.TRIB.EXFONT	49.38	3.74	48.57	0.98	12.98
VARIAÇÃO DE BENS	44.70	2.78	51.33	1.15	18.45

algum valor positivo para esta variável. Como esperado os ganhos mais elevados estão associados às variáveis que foram incluídas nos procedimentos de estratificação e alocação.

Os ganhos G_4 e G_5 revelam aspectos interessantes associados ao cruzamento da estratificação por Renda Tributável e Condição de Especial. Por eles observa-se que o uso da estratificação, por Renda Tributável traz ganhos expressivos apenas para as variáveis positivamente correlacionadas com ela, não influindo ou mesmo causando perdas, para as demais variáveis.

A estratificação por Condição de Especial tem o efeito contrário, ou seja, só é significativa para as variáveis não correlacionadas com a Renda Tributável. As conclusões dizem respeito ao ganho adicional das estratificações, obviamente. Pode-se concluir, assim, que a estratificação cruzada produziu exatamente o que se esperava dela, ou seja a redução simultânea das variâncias dos estimadores dos totais de variáveis correlacionadas e não-correlacionadas com a Renda Tributável.

5 – CONCLUSÕES

O uso de técnicas simples da teoria de amostragem permite uma redução considerável do volume de dados necessários para fins de análise de grandes arquivos. Na presente aplicação, esta redução é particularmente crítica já que o propósito fundamental é o de simular efeitos decorrentes de alterações introduzidas na legislação tributária.

Tais procedimentos consistem, essencialmente, em refazer a declaração de rendimentos de cada contribuinte à luz da nova legislação a ser simulada.

A utilização da totalidade dos registros que compõem o universo de declarantes do IRPF ($\sim 8.3 \times 10^6$) inviabilizaria o processo, tendo-se em vista o uso repetitivo do Sistema Simulador, que é um sistema típico de apoio a tomada de decisão.

A autoridade tributária normalmente está interessada em simular um conjunto de alterações a serem introduzidas na legislação do IRPF.

O caminho adotado é simular cada alteração em separado, medindo o seu efeito marginal através da análise de um conjunto de indicadores adequados para cada alteração, e em seguida avaliar o impacto da adoção simultânea de todas as alterações (efeito conjunto).

Assim, por exemplo, durante a elaboração do Pacote Fiscal-87 (Plano Bresser), foram realizadas dezenas de simulações até se chegar à solução final adotada pelo governo.

O emprego de uma amostra reduzida de declarantes ($\sim$ 1% do total) permite uma rapidez de processamento essencial para os objetivos de simulação, que é o de fornecer respostas rápidas e precisas para o órgão gestor do tributo.

Evidentemente, o uso de amostras de grandes arquivos possui outros atrativos. Entre outros igualmente relevantes, poderíamos citar: gerar estatísticas a baixo custo e a curto prazo, possibilitar a substituição de apurações especiais por uma alternativa mais rápida, mais barata e suficientemente precisa para um grande número de aplicações, oferecer uma excelente alternativa para realização de testes de programas, etc.

O SERPRO possui hoje um conjunto de amostras estratificadas do IRPF (de 1975 até 1987) que se constitui em um acervo importante para estudos na área de política econômica, política fiscal, etc.

Alternativas para os procedimentos de estratificação, alocação e estimação das principais características vêm sendo pesquisadas, visando à redução considerável do tamanho da amostra, sem perda significativa de precisão.

Dentro desta linha de estudo, pretende-se futuramente dimensionar e coletar uma amostra passível de utilização em microcomputador, o que abriria uma nova perspectiva de trabalho para pessoas envolvidas com análise de legislação tributária.

BIBLIOGRAFIA

COCHRAN, W. *Sampling Techniques*. 3th ed. New York : John Wiley, 1977.

DALENIUS, T., HODGES, J.L. Minimum variance stratification. *Journal of the American Statistical Association*, v. 54 n. 285, p. 88–101, 1959.

XAVIER, A. et al. Simulador de legislações para o imposto de renda pessoa física (IRPF). In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 2, 1976, Campinas, SP. *Atas...* Campinas, SP : Simpósio Nacional de Probabilidade e Estatística, 1976.

RESUMO

A cada ano, o Ministério da Fazenda, através da Secretaria da Receita Federal – SRF, altera alguns dos parâmetros que definem a Legislação do Imposto de Renda Pessoa Física – IRPF.

Alterações na legislação do IRPF geram, portanto, a necessidade de se estimar previamente os efeitos delas decorrentes, uma vez que a adoção de uma nova legislação promoverá reflexos do ponto de vista econômico, fiscal e administrativo.

De forma a atender às necessidades da SRF, o Serviço Federal de Processamento de Dados – SERPRO desenvolveu o Sistema Simulador de Legislações, que vem sendo usado ao longo dos últimos 12 anos como uma ferramenta importante no apoio à tomada de decisão na área tributária. Este sistema utiliza diversas técnicas de amostragem tais como estratificação e alocação ótima de amostras.

Este trabalho ilustra o procedimento padrão adotado para estratificação da população de declarantes do IRPF e alocação, seleção e validação da amostra a ser utilizada no Sistema Simulador.

ABSTRACT

Every year, the income revenue service (IRS) modifies certain parameters that define the income tax legislation. The possible consequences of these modifications must be estimated in advance, because of its great economical, fiscal and administrative impacts.

In order to atend the needs of the IRS, the Federal Service of Data Processing has developped the Legislation Simulator System. This system has been in use for the last twelve years. It makes use of several statistical tools, with an emphasis in sempling technics.

This article describes the standard procedures employed in selecting a sample of tax-payers from the files of the IRS, for use in the Simulator System.

DETERMINAÇÃO DE SEMELHANÇAS REGIONAIS: UMA METODOLOGIA UTILIZANDO ANÁLISE DAS COMPONENTES PRINCIPAIS

Lucia Silva Kubrusly*
e
Deborah Roditi*

1 – INTRODUÇÃO

Sejam $r_s(s = 1, 2, \cdots, S)$ as regiões sobre as quais desejamos estabelecer uma comparação segundo certo critério escolhido *a priori*. No presente trabalho o critério escolhido será representado por um conjunto de variáveis $X_k(k = 1, 2, \ldots, m)$ observadas para todas as regiões r_s. O objetivo da análise apresentada é verificar se as regiões r_s são semelhantes ou não segundo o grupo de variáveis escolhidas. Caso não se verifique a semelhança para todas a regiões r_s, é interessante tentar determinar subconjuntos de regiões semelhantes, ainda segundo o mesmo critério.

2 – CARACTERIZAÇÃO DAS REGIÕES

A fim de determinar semelhanças e diferenças entre as regiões analisadas, é necessário caracterizar cada região. Neste trabalho essa caracterização foi feita a partir

*Pesquisadoras do Laboratório Nacional de Computação Científica - LNCC/CNPq.

das componentes principais extraídas do conjunto de variáveis escolhidas, para cada região, como pode ser visto em (Krzanouskik, 1984, e em Stemmelen, 1977).

Portanto, sendo $X^s \in R^m$ o vetor das variáveis observadas para cada região $r_s, s = 1, 2, \cdots, S$ e Σ^s a matriz de covariância de X^s, as componentes principais são dadas por (Johnson e Wichern - 1982).

$$C_i^s = (c^s)^t X^s \text{ tal que}$$
$$var(C_i^s) = (c^s)^t \Sigma^s c^s = \max,$$

onde $(c^s)^t$ é o vetor transposto de C^s

$$(c^s)^t c^s = 1$$
$$corr(C_i^s, C_{i'}^s) = 0 \quad \forall i \neq i'$$
$$i = 1, 2, \ldots, p < m$$

onde p é o número de componentes principais que descrevem no mínimo P% da variância total dos dados iniciais.

Para cada região s, são extraídas p componentes principais, as quais serão utilizadas para verificar as semelhanças e diferenças entre as regiões consideradas. É portanto necessário escolher parâmetros que identifiquem essas componentes para que tal verificação seja possível. Nesse trabalho essa caracterização será abordada conforme (Gouvêa e Kubrusly 1985).

Definição 1. Contribuição da variável X_k^s para a componente C_i^s.

Sendo $C_i^s = \sum_{k=1}^m c_{ik}^s X_i^s$ então $(c_{ik}^s)^2$ chama-se contribuição da variável X_k^s na composição da componente C_i^s.

Definição 2. Ordem da componente.

A ordem i $(i = 1, 2, \cdots, p)$ da componente C_i^s é dada pela relação de ordem decrescente das variâncias das componentes extraídas na região s (supondo que exista uma ordem estrita para essas variâncias, isto é, $\lambda_1 > \lambda_2 > \cdots > \lambda_p$).

Com essas duas definições pode-se caracterizar cada componente de duas formas. Ou pela sua ordem i, ou pelas variáveis que contribuem mais fortemente para sua composição.

3 – CRITÉRIOS DE SEMELHANÇA

A fim de estabelecer critérios que possibilitem verificar a semelhança entre regiões, partiremos de uma idéia de semelhança baseada na proporcionalidade das variáveis tomadas para diferentes regiões.

Definição 3. Semelhança máxima.

Considere um conjunto de m variáveis X_k $(k = 1, 2, \ldots, m)$ tomadas sobre as regiões $r_s (s = 1, 2, \ldots, S)$. Dizemos que as regiões apresentam semelhança máxima

quando para todo $s \neq s'$ temos:

$$X_k^s = \alpha_k^{s'} X_k^{s'} \quad \forall k, \quad \text{com} \quad \alpha_k^{s'} \in \mathbf{R}$$

A partir dessa definição, vê-se imediatamente que, se as componentes principais são extraídas da matriz de correlação dos dados, regiões com semelhança máxima fornecerão componentes iguais pela ordem. Isto é, se a regiões s e s' apresentam semelhança máxima e R^s e $R^{s'}$ são as respectivas matrizes de correlação, então

$$X_k^s = \alpha_k^{s'} X_k^{s'} \Rightarrow R^s = R^{s'} \Rightarrow C_i^s = C_i^{s'} \quad \text{e} \quad \lambda_i^s = \lambda_i^{s'}$$

já que as componentes principais C_i^s e suas variâncias λ_i^s são os autovetores e autovalores da matriz de correlação dos dados. Para o caso das componentes serem extraídas a partir da matriz de covariância, a definição de semelhança máxima pode ser mantida alterando-se apenas $X_k^s = \alpha^{s'} X_k^{s'}, \alpha^{s'} \in \mathbf{R}$ (veja Kubrusly, 1985).

Os critérios de semelhança propostos a seguir são na realidade um relaxamento da condição de semelhança máxima entre regiões. Isto é, ao invés da igualdade das componentes de mesma ordem i, os critérios estabelecidos definem condições de "semelhança" entre essas componentes.

Definição 4. Critérios de semelhança das componentes segundo seus coeficientes.

Seja o conjunto de componentes principais $C_i^s (i = 1, 2, \cdots, p;\ s = 1, 2, \cdots, S)$. Para cada componente C_i^s considere o conjunto V_i^s formado pelas variáveis X_{k*}^s que contribuem fortemente para a componente C_i^s, isto é:

$$V_i^s = \{X_{k^*}^s\} \quad \text{onde} \quad k^* \subset k = \{1, 2, \ldots, m\},$$

é o conjunto dos índices das variáveis cujos coeficientes $(c_{i_{k*}}^s)^2 \geq (c_{min})^2$, onde $(c_{min})^2$ é o mínimo valor que $(c_{i_k}^s)^2$ pode assumir para que a variável X_k^s tenha uma contribuição significativa para C_i^s.

Para cada ordem i, as componentes $C_i^s (i$ fixo, $s = 1, 2, \ldots, S)$ são consideradas semelhantes se $V_i^1 = V_i^2 = \cdots = V_i^S$.

O segundo critério de semelhança proposto procura verificar duas condições: a semelhança entre as componentes de mesma ordem i e a diferença entre componentes de ordens distintas. Essa diferença é justificada pela propriedade de não correlação entre as componentes de ordem distintas (veja Johnson e Wichern, 1982).

Assim, procura-se verificar:

itemi) $C_i^s (i$ fixo; $s = 1, 2, \ldots, S)$ são semelhantes;

ii) C_i^s, e $C_i^s (s = 1, 2, \ldots, S;\ i, i' = 1, 2, \ldots, p;\ i \neq i')$ são diferentes.

O critério proposto a seguir, ao invés de examinar cada componente C_i^s separadamente, considera o conjunto de todas as componentes $C_i^s (i = 1, 2, \cdots, p;\ s = 1, 2, \cdots, S)$ e, a partir deste, verifica se as duas condições são ou não satisfeitas. Para isso as componentes C_i^s são consideradas como um conjunto de pS pontos no espaço R^m formando uma matriz de dados $(m \times pS)$. Desses novos dados serão extraídas as componentes principais $\mathbf{K}_j$.

$$\mathbf{K}_j = \sum_{i=1}^{p}\sum_{s=1}^{S} k_{ij}^s C_i^s$$

$$var(\mathbf{K}) = max$$

$$cor(\mathbf{K}_j, \mathbf{K}_{j'}) = 0 \quad \forall j \neq j'$$

Tomando as $q < m$ primeiras componentes $\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_q$, pode-se representar cada componente C_i^s como um ponto no espaço $\mathbf{R}^q$ cujo sistema de referências é formado pelas componentes $\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_q$. É portanto nesse espaço IR^q que as "semelhanças" e "diferenças" entre as componentes C_i^s serão verificadas.

Definição 5. Esfera de ordem i.

Sejam $C_i^s (i = 1, 2, \ldots, p;\ s = 1, 2, \ldots, S)$, pS pontos no espaço IR^q. Considere o ponto

$$0_i = \frac{1}{2}\sum_{s=1}^{s} C_i^s$$

média dos pontos $C_i^s (i$ fixo, $s = 1, 2, \ldots, S)$ e seja

$$r_i = \max_s \{d(0_i, C_i^s\},$$

onde $d(0_i, C_i^s)$ é a distância entre 0_i e C_i^s. Chama-se de esfera de ordem i, E_i, a esfera de centro 0_i e raio r_i, isto é,

$$E_i = \{x \in IR^q; d(x, 0_i) \leq r_i\}$$

Propriedades da esfera E_i

i) E_i contém todos os pontos $C_i^s (i$ fixo, $s = 1, 2, \ldots, S)$; e

ii) E_i é a menor esfera de centro 0_i que contém todos os pontos $C_i^s (i$ fixo, $s = 1, 2, \ldots, S)$.

Definição 6. Critérios de semelhança das esferas disjuntas.

Sejam as componentes principais C_i^s e considere o espaço $\mathbf{R}^q$ definido pelas componentes $\mathbf{K}_j$, no qual se constroem as esferas $E_i (i = 1, 2, \ldots, p)$ definidas anteriormente. Consideram-se semelhantes as componentes C_i^s se as esferas E_i são disjuntas para todo i.

No critério definido acima procura-se formar grupos de componentes de mesma ordem i. Se para cada ordem i for possível formar grupos diferentes, então as componentes C_i^s são semelhantes, e as regiões $r_s (s = 1, 2, \ldots, S)$ as quais são caracterizadas por essas componentes são consideradas semelhantes entre si.

4 – APLICAÇÃO

A seguir será apresentada uma aplicação da caracterização de regiões utilizando-se a análise das componentes principais. Foram utilizados os dados da ENDEF, 1972,

relativos à posse de aparelhos domésticos, observados para diversas regiões brasileiras. A seguir serão apresentadas as variáveis e regiões escolhidas.

4.1 – Descrição das Variáveis e Regiões

Os dados utilizados indicam a proporção das famílias que possuem bens duráveis por classe de despesa monetária global, segundo o tipo de bem. Os bens duráveis escolhidos nesse trabalho são os aparelhos domésticos relacionados abaixo:

Variáveis

REFRI - refrigerador LAVAR - máquina de lavar COSTURA - máquina de costura ENCERA - enceradeira ASPO - aspirador de pó FOGAS - fogão a gás FOLENHA - fogão a lenha FERELET - ferro elétrico VENT - ventilador LIQUI - liquidificador BATILET - batedeira elétrica

Regiões

Foi utilizada a classificação regional adotada na Pesquisa Nacional por Amostra de Domicílios-PNAD, subdividida em regiões metropolitana, urbana não-metropolitana e rural, totalizando 22 regiões brasileiras, descritas a seguir:

REGIÃO I - Estado do Rio de Janeiro

RJM - Rio de Janeiro metropolitana

RJURBE - Rio de Janeiro urbana não-metropolitana

RJRURAL - Rio de Janeiro rural

REGIÃO II - Estado de São Paulo

SPM - São Paulo metropolitana

SPURBE - São Paulo urbana não-metropolitana

SPRURAL - São Paulo rural

REGIÃO III - Paraná, Santa Catarina e Rio Grande do Sul

CURIT - Curitiba

PORTAL - Porto Alegre

SULURBE - Região III urbana não-metropolitana

SULRURAL - Região III rural

REGIÃO IV - Minas Gerais e Espírito Santo

BH - Belo Horizonte

IVURBE - Região IV urbana não-metropolitana

IVRURAL - Região IV rural

REGIÃO V - Bahia, Sergipe, Alagoas, Pernambuco, Paraíba, Rio Grande do Norte, Ceará, Piauí e Maranhão

SALVA - Salvador

RECI - Recife

FORTAL - Fortaleza

NEURBE - Região V urbana não-metropolitana

NERURAL - Região V rural

REGIÃO VI - Brasília (DF)

REGIÃO VII - Rondônia, Acre, Amazonas, Roraima, Pará, Amapá, Mato Grosso, Mato Grosso do Sul e Goiás

BELEM - Belém

ROURBE - Rondônia, Acre, Amazonas, Roraima, Pará, Amapá (urbana não-metropolitana)

GOURBE - Mato Grosso, Mato Grosso do Sul, Goiás (urbana não-metropolitana)

4.2 – Análise de Semelhança Regional

Para caracterização e comparação das regiões segundo o critério estabelecido pelo conjunto de variáveis relativas à posse de aparelhos domésticos, foram dados os seguintes passos:

i) caracterização de cada região através das suas componentes principais;

ii) comparação das componentes principais utilizando os critérios definidos na seção 3; e

iii) formação dos grupos de regiões cujas componentes principais são semelhantes, obtendo-se dessa forma grupos de regiões homogêneas segundo o critério escolhido.

Na primeira etapa foram extraídas as componentes principais para cada uma das 22 regiões escolhidas. Foram sempre mantidas duas componentes principais para cada região, correspondendo a no mínimo 80,59% da variância total.

Apenas observando os coeficientes das componentes principais relativas às diversas regiões, pode-se notar que existem diferenças marcantes entre elas. Apenas, como exemplo, ressaltamos as regiões Rio de Janeiro metropolitana, São Paulo metropolitana, Rio de Janeiro urbana e São Paulo urbana, as quais apresentam suas componentes de ordem 1 principalmente descritas pelas variáveis REFRI, COSTURA, ENCERA, FOGAS, FERELET e LIQUI, enquanto que para outras regiões, tais como Belo Horizonte, as variáveis ASPO e BATILET são dominantes na primeira componente, enquanto que as variáveis FOGAS e FERELET aparecem na composição da segunda componente principal.

Muitas outras diferenças e semelhanças podem ser observadas apenas analisando-se os coeficientes de cada componente para cada região. Nesse trabalho optamos pela utilização dos critérios de semelhança definidos na seção anterior. Primeiramente será aplicado o critério das esferas disjuntas para todas as regiões. Isto é, serão tomadas as primeiras e segundas componentes (C_1^s, C_2^s) de todas as regiões analisadas, e a partir destas serão extraídas as componentes principais $\mathbf{K}_1$ e $\mathbf{K}_2$, e, finalmente serão construídas as esferas E_1 (contendo todas as componentes C_1^s) e E_2 (contendo todas as componentes C_2^s). Se as esferas assim construídas forem disjuntas, então pode-se concluir que todas as regiões são semelhantes segundo o critério definido pelas variáveis

"posse de aparelhos domésticos".

A Figura 1 mostra os pontos C_1^s e C_2^s no plano $\mathbf{K}_1 \times \mathbf{K}_2$. As esferas E_1 e E_2 não são absolutamente disjuntas, no entanto alguns aglomerados de pontos podem ser observados, dando uma primeira informação sobre regiões possivelmente semelhantes.

À direita da Figura 1, observa-se um grande ajuntamento de pontos relativos à componente C_1, e alguns pontos relativos à C_2. Do lado esquerdo se dá o inverso. Uma primeira aproximação da formação de grupos homogêneos pode ser obtida procurando-se grupos de regiões cujas componentes C_1 estão juntas e cujas componentes C_2 também estão juntas. Destaca-se imediatamente o grupo FORTAL,SALVA,RECI e NEURBE, com as componentes C_1 localizadas à direita no alto e as componentes C_2 no centro embaixo da Figura 1. Outro possível grupo homogêneo seria o formado por RJM, SPM, RJU, SPU e PORTAL, com as componentes C_1 no alto à esquerda e as componentes C_2 à direita, próximo do eixo de $\mathbf{K}_1$. As demais regiões não podem ser grupadas apenas observando-se a Figura 1. A seguir, será feita uma tentativa de formar grupos, utilizando-se os critérios definidos na seção 3.

4.3 – Formação de Grupos Homogêneos de Regiões

Com a utilização do critério dos coeficientes das componentes C_i^s apresentado na Seção 3, é possível identificar quais as regiões que apresentam o mesmo conjunto de variáveis dominantes para componentes de mesma ordem. São os conjuntos V_1^s e V_2^s. Utilizando-se esse critério é possível formar os seguintes grupos:

Grupo 1: RJM, SPM, RJURB, SPURB e FORTAL

V_1 = REFRI,COSTURA,ENCERA,FOGAS,FERELET,LIQUI
V_2 = LAVAR,ASPO,FOLENHA,BATILET

Grupo 2: CURIT e SULURB

V_1 = REFRI,LAVAR,ENCERA,VENT,LIQUI
V_2 = COSTURA,FOGAS,FOLENHA,FERELET

Grupo 3: RJRU, SPRU e IVURB

V_1 = REFRI,LAVAR,ENCERA,ASPO,VENT,LIQUI
V_2 = COSTURA,FOGAS,FERELET

Grupo 4: FORTAL, SALVA, RECI e NEURBE

V_1 = REFRI,LAVAR,ENCERA,ASPO,FOGAS,FERELET,VENT,LIQUI, BATILET
V_2 = FOLENHA

Grupo 5: BH e DF

V_1 = REFRI,LAVAR,COSTURA,ENCERA,ASPO,VENT,LIQUI,BATILET

V_2 = FOGAS,FERELET

Grupo 6: BELÉM, GOURBE, ROURBE, IVRURAL e NERURAL

V_1 = ENCERA,ASPO,FOLENHA,VENT,LIQUI,BATILET
V_2 = COSTURA,FOGAS,FERELET

A fim de verificar se os grupos formados acima são realmente homogêneos, será aplicado o critério das esferas disjuntas para cada um deles. As Figuras 2 a 7 mostram a configuração dos pontos C_1 e C_2 para cada grupo. Observa-se que pelo critério adotado todos os grupos são considerados homogêneos, mas é possível avaliar o grau de homogeneidade dos grupos, pelo raio da esfera E_1 obtida conforme a definição 5 (foi escolhida E_1 porque as componentes que a compõem descrevem uma maior proporção da variância total dos dados).

Sendo assim, os grupos 3, 4 e 5 os mais homogêneos e o grupo 2 é o menos homogêneo. Vale a pena chamar atenção para o fato de que a Região Sul (Porto Alegre, Curitiba, Sul urbana e Sul rural) mostrou-se bastante heterogênea nesta análise. Porto Alegre foi classificada no Grupo 1 (junto com Rio de Janeiro e São Paulo, e a Região Sul rural não pode ser classificada em nenhum grupo. Outro aspecto interessante que a análise segundo o critério das esferas disjuntas revela é uma "divisão" do Grupo 6, indicando dois aglomerados distintos na esfera E_1: um formado pelas regiões IVRURAL e NERURAL, e outro formado por BELÉM, ROURBE e GOURBE. Pode-se dizer neste caso que o Grupo 6 ficou subdividido em um Grupo 6-urbano, e outro 6-rural. Outro aspecto observável pelo critério das esferas, é o relativo afastamento da região PORTAL no Grupo 1 (Figura 2). Nota-se que o grupo seria muito mais homogêneo sem a presença desta região, mas por outro lado, PORTAL não pode ser melhor alocada em nenhum dos grupos restantes.

5 – CONCLUSÕES

A metodologia apresentada neste trabalho, consistindo da utilização dos critérios de semelhança definidos na Seção 3, possibilitou, em primeiro lugar, pelo critério das esferas disjuntas, a constatação da não homogeneidade das 22 regiões escolhidas; em segundo lugar, através do critério dos coeficientes, a formação de grupos homogêneos de regiões. E finalmente, mais uma vez pelo critério das esferas disjuntas, foi possível a confirmação dos grupos homogêneos, e uma análise do grau de homogeneidade dos mesmos.

É importante lembrar que a presente análise foi facilitada pela representação bidimensional das esferas no espaço de $\mathbf{K}_1$ e $\mathbf{K}_2$. Isso foi possível devido à grande proporção da variância contida nas duas primeiras componentes. Evidentemente os critérios definidos na Seção 3 são válidos qualquer que seja o número de componentes mantidas para análise (Kubrusly e Gouvêa 1988), mas a forte estrutura de correlação das variáveis escolhidas tornou possível a visualização dos resultados no plano $\mathbf{K}_1 \times \mathbf{K}_2$.

FIGURA 1

AS ESFERAS E_1 e E_2 PARA AS 22 REGIÕES

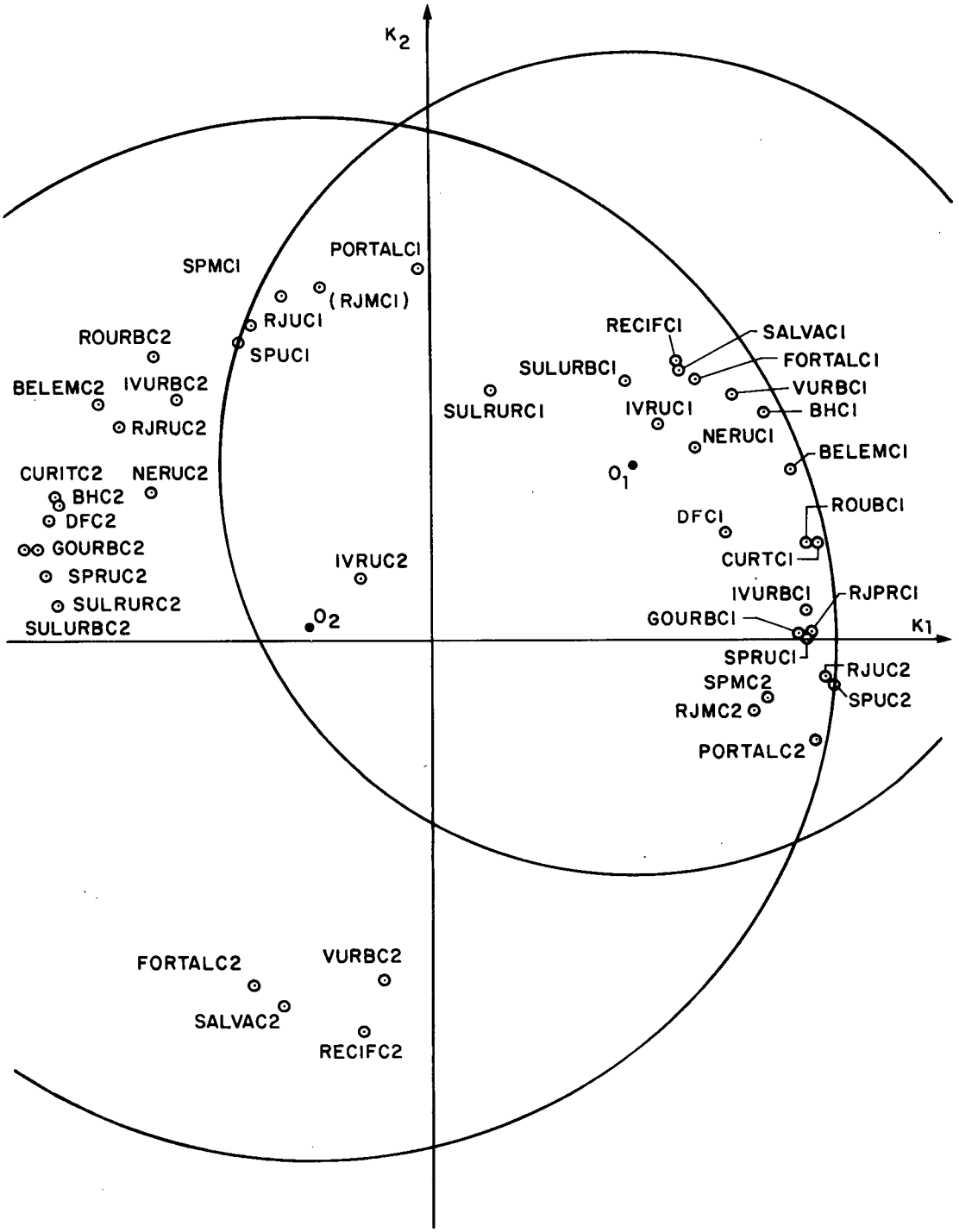


FIGURA 2

GRUPO 1: RJMETRO, SPMETRO, RJURB, SPURB, PORTAL.

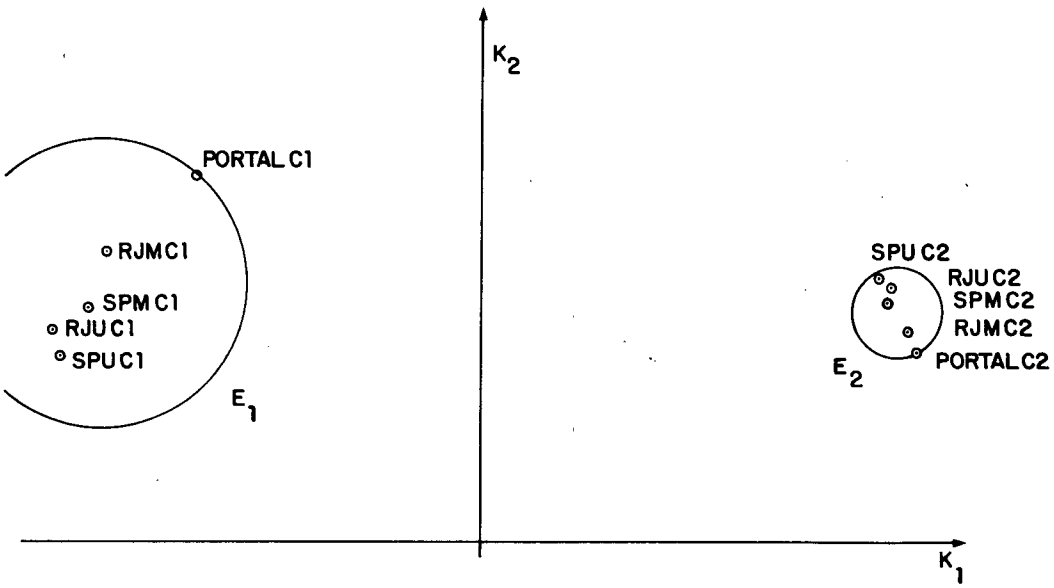


FIGURA 3

GRUPO 2: CURIT, SULURB.

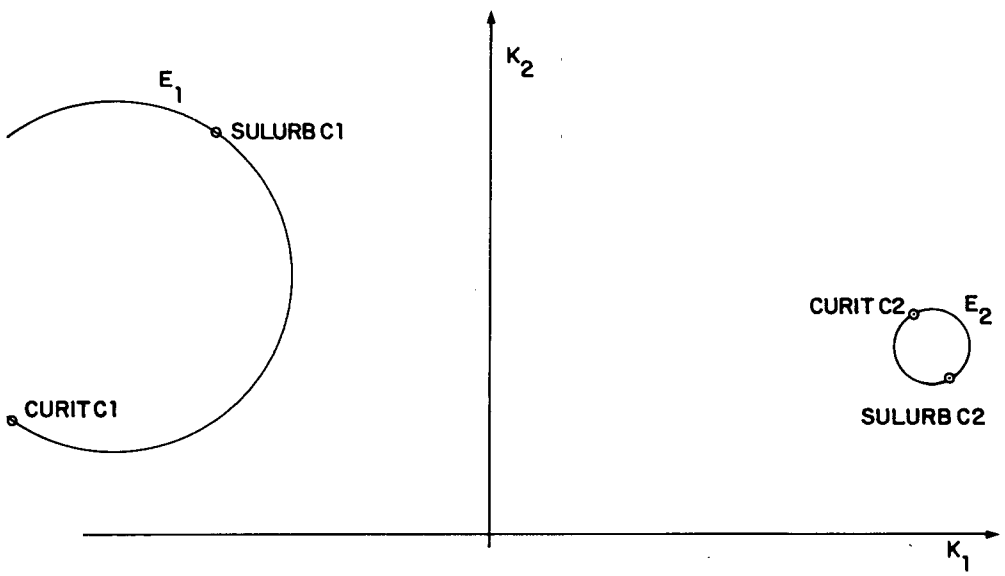


FIGURA 4

GRUPO 3: RJRURAL, SPRURAL, IVURB.

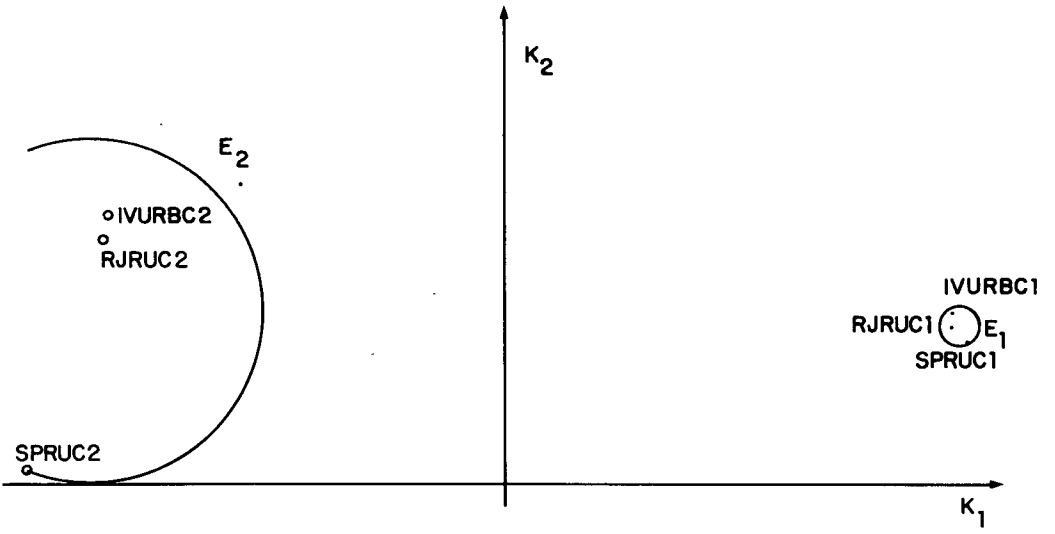


FIGURA 5

GRUPO 4: FORTAL, RECI, SALVA, NEURB.

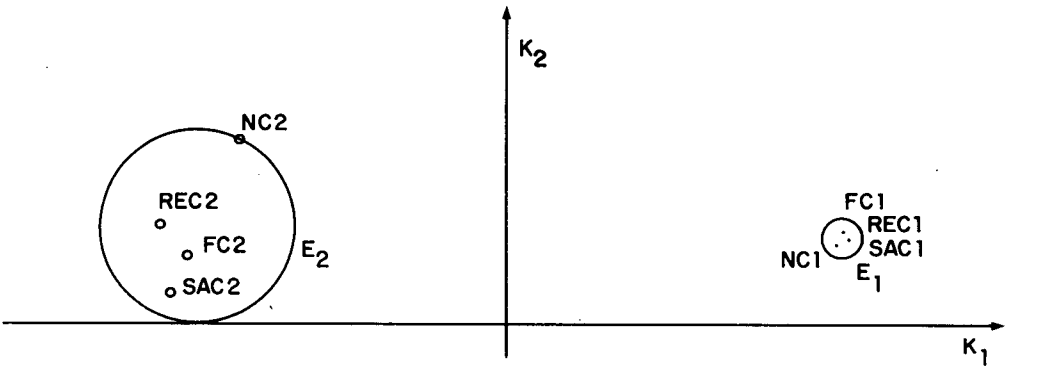


FIGURA 6

GRUPO 5: BH, DF.

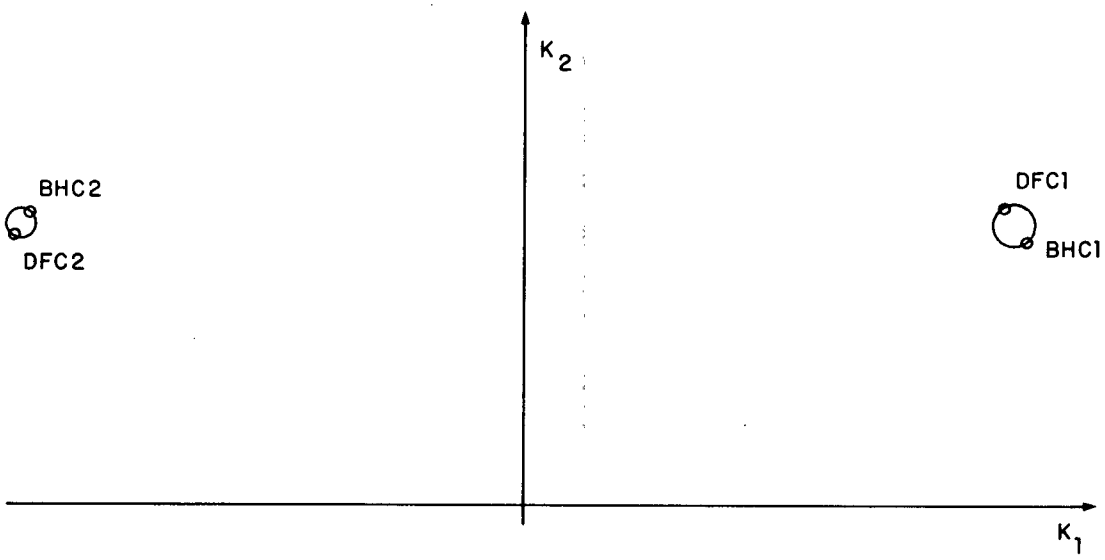
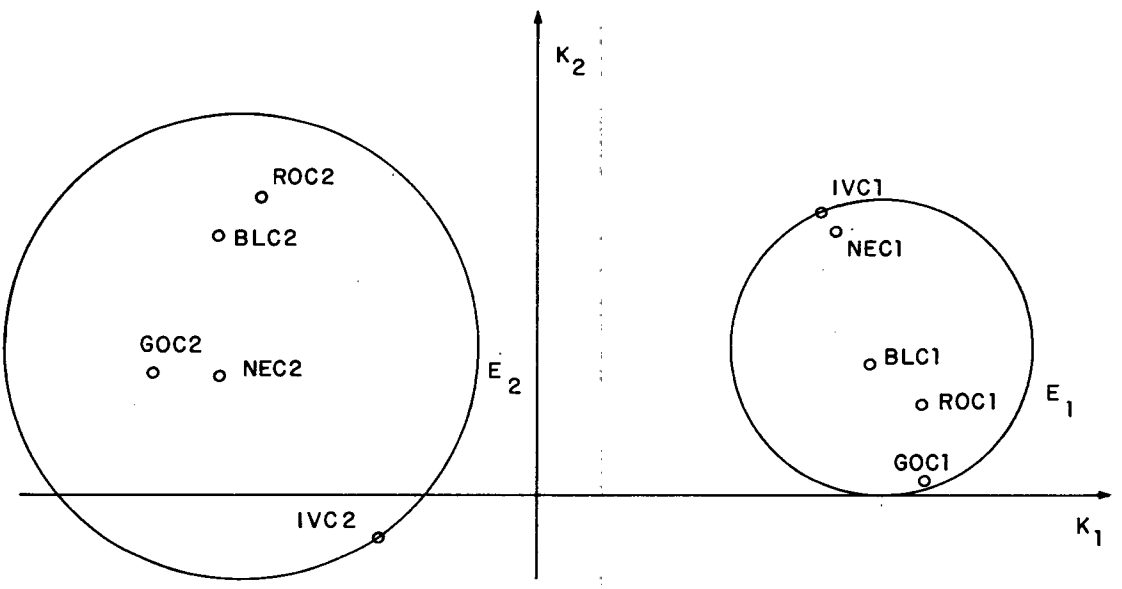


FIGURA 7

GRUPO 6: BELEM, GOURB, ROURB, IVRURAL, NERURAL.



BIBLIOGRAFIA

ESTUDO Nacional de Despesa Familiar - ENDEF. Rio de Janeiro : IBGE, 1972.

GOUVÊA, V.H.C. *Estudo da estabilidade de fatores em análise fatorial de correspondência para previsão de matrizes.* Niterói, RJ : UFF, 1980. Tese para Prof. Titular de Estatística da Universidade Federal Fluminense.

JOHNSON, R. A; WICHERN D.W. *Applied multivariate statistical analysis.* Englewood Cliffs, New Jersey : Prentice-Hall, 1982.

KRZANOUSKY, W.J. Principal component analysis in the presence of group structure. *Applied Statistics*, v. 33, n. 2, p. 164-68, 1984.

KUBRUSLY, L.S. *Critérios de estabilidade para o problema de previsão de matrizes.* Rio de Janeiro : COPPE/UFRJ, 1985. - Tese (Doutorado em Engenharia de Sistemas).

KUBRUSLY, L.S. e GOUVÊA, V.H.C. Análise da estabilidade no tempo das matrizes do balanço energético. (1976-1980). *Revista de Econometria*, v. 8, n. 2, p. 93–107, nov. 1988.

STEMMELEN, E. *Tableaux d'echange: description et prevision.* Paris : Université de Paris VI, 1977. (Cahier du Bureau Universitaire de Recherche Operationelle, n. 28).

RESUMO

O objetivo desse trabalho é apresentar uma metodologia que permita verificar se um conjunto de regiões é homogêneo, isto é, se as regiões que o compõem são semelhantes entre si. A metodologia proposta pode ser dividida em três fases. A primeira trata da caracterização das regiões pelas suas componentes principais. Na segunda fase são definidos parâmetros de identificação que possibilitam comparar as componentes relativas a diferentes regiões. E finalmente, na terceira fase, são estabelecidos critérios de semelhança regional.

Uma aplicação dessa metodologia é apresentada utilizando-se dados da ENDEF relativos à posse de aparelhos domésticos observados para as regiões PNAD.

ABSTRACT

This paper presents an analysis of regional similarities. The proposed methodology can be splited in three steps. In the first step, the regions are characterized by its principal components. In the second step, identification parameters are defined and used for compairing the principal components. Finally, in the third step, regional similarity criteria are established.

An application of this methodology is presented, making use of ENDEF data related to the possession of domestic appliance utilities.

UMA APLICAÇÃO DA FUNÇÃO DE GOMPERTZ NA ANÁLISE E NA PROJEÇÃO DE DOMICÍLIOS POR CLASSES DE TAMANHO

Ricardo F. Neupert*

1 – INTRODUÇÃO

A necessidade de projeções domiciliares tem aumentado de forma considerável nos últimos anos, em conseqüência da crescente importância das atividades de planejamento econômico e social. Essas projeções são fundamentais para determinar a demanda futura de habitação, serviços de utilidade pública, assim como a produção e distribuição de bens de consumo duráveis, como eletrodomésticos, móveis e carros particulares, cujo consumo é mais familiar ou domiciliar, do que individual. A abrangência dessas previsões implica não só o conhecimento do número futuro de domicílios, mas também sua composição em função do total de indivíduos que abrigam, já que para o planejamento da construção de moradias trabalha-se tanto em termos da quantidade quanto do tamanho das unidades. Esses cálculos são ainda importantes na estimação do consumo futuro de serviços como energia elétrica doméstica, cujas projeções dependem também do número de unidades domiciliares e do número de residentes por domicílio.

Para a elaboração desses estudos, uma dificuldade que se apresenta diz respeito a técnicas apropriadas. A principal fonte de técnicas de projeções domiciliares, desde

*Consultor do Projeto de Planejamento e Avaliação Social - PNUD/OIT/IPEA.

sua publicação em 1973 pelas Nações Unidas, tem sido o Manual VII "Methods of Projecting Households and Families". Este manual recomenda o uso de técnicas baseadas em taxas específicas de chefes de domicílio por sexo e idade (e, se possível, também por estado conjugal), dada a disponibilidade e a qualidade da informação censitária na maioria dos países. Além disso, empregam como base projeções populacionais por sexo e idade, refletindo as mudanças na composição da população, as que, certamente, afetam o número de domicílios. Permitem ainda separar o efeito das alterações populacionais daquelas ocorridas na tendência à privacidade residencial no número futuro de domicílios (United Nations, 1973). Porém, essas técnicas possibilitam apenas a projeção do número total de domicílios, segundo algumas características dos seus chefes e não por classes de tamanho. O método mais adequado a esta última informação, também apresentado no Manual VII das Nações Unidas, baseia-se nos modelos de Brown-Glass-Davidson. Entretanto, durante os últimos anos, importantes avanços no desenvolvimento de modelos de formação de famílias tornaram possíveis a obtenção de informações detalhadas sobre a estrutura domiciliar (Bongaarts, 1981, Menken, 1985; Watkins, Et. Al., 1984; Yi, 1986). Estes modelos têm uma aplicação limitada na projeção de domicílios em muitos países, seja por se utilizarem de pressupostos nem sempre compatíveis com a realidade ou, principalmente, porque o tipo de informação requerida não se encontra freqüentemente disponível. Dados detalhados e confiáveis sobre fecundidade, mortalidade, nupcialidade, divórcio e parturição são, em geral, de difícil acesso na maioria dos países subdesenvolvidos.

O objetivo do presente trabalho é propor uma técnica, baseada no uso da função de Gompertz, para representar o número de domicílios por classes de tamanho. Por extensão, também poderia ser utilizada para desagregar, por classes de tamanho, uma projeção disponível do número total de domicílios. Esta técnica não considera diretamente os aspectos dinâmicos do ciclo de vida familiar, isto é, formação, crescimento, contração e dissolução de domicílios e famílias. Apresenta, porém, uma vantagem metodológica: utiliza informações normalmente disponíveis na maioria das publicações censitárias.

2 – DESENVOLVIMENTO DO MODELO

A função de Gompertz tem sido fundamental ao estudo da mortalidade e da fecundidade (Brass, 1974; Gompertz, 1825; Martin, 1967; Wunsch, 1966). Constitui, também, um instrumento bastante adequado à projeção dessas variáveis demográficas (CELADE, 1983; Farid, 1973; Romaniuk e Tanny, 1969).

A forma geral da função é:

$$Y = kA^{B^x} \tag{2.1}$$

onde k é uma assíntota e A e B são os parâmetros da função.

Neste trabalho, propomos a possibilidade de utilização desta função na apresentação da distribuição relativa acumulada de domicílios por classes de tamanho. Seja N_x o número de domicílios acumulados por classes, x as classes (1, 2, 3, etc.) e T o número total de domicílios:

$$N_x = TA^{B^x}$$

A distribuição relativa acumulada tem a seguinte forma:

$$N_x/T = A^{B^x}$$

ou

$$D_x = A^{B^x} \tag{2.2}$$

Onde $D_x = N_x/T$ e $0 < A < 1 \quad 0 < B < 1 \quad 0 < D$

Aplicando logaritmo natural em (2), tem-se:

$$\log(-\log D_x) = x \log B + \log(-\log A) \tag{2.3}$$

onde log denota Logaritmo Neuperiano.

O que pode ser considerado uma reta da forma:

$$y_x = \alpha_0 + \beta_0 x \tag{2.4}$$

Onde:

$$y_x = \log(-\log D_x), \quad \alpha_0 = \log(-\log A),$$
$$\beta_0 = \log B, \quad -\infty < \alpha_0 < \infty \quad \text{e} \quad 0 < \beta < \infty$$

Considere-se, agora, uma distribuição padrão da proporção de domicílios por classe de tamanho:

$$y_x^p = \alpha_p + \beta_p x \tag{2.5}$$

Onde:

$$y_x^p = \log(-\log D_x^p)$$

Se y_x e y_x^p possuem comportamento linear com respeito à classe de tamanho x, também haverá uma relação linear entre eles, posto que de (5) obtém-se:

$$x = 1/\beta_p\, y_x^p - \alpha_p/\beta_p$$

Que, substituindo em (4), resulta em:

$$y_x = \alpha_0 - \alpha_p \beta_0/\beta_p + \beta_0/\beta_p\, y_x^p$$

E chamando:

$$\alpha = \alpha_p - \alpha_p \beta_0/\beta_p \quad \text{e} \quad \beta = \beta_0/\beta_p$$

Tem-se que:

$$y_x = \alpha + \beta y_x^p \tag{2.6}$$

Os parâmetros α e β da equação (6) são calculados ajustando-se a uma reta os valores obtidos da linearização das freqüências relativas acumuladas da distribuição observada e padrão (y_x e y_x^p). A técnica utilizada para tal fim pode ser a das médias[1] .

Note-se que se $y_x = y_x^p$ tem-se que $\alpha = 0$ e $\beta = 1$

α e β são os parâmetros que diferenciam a distribuição observada dos domicílios, por classes de tamanho, da distribuição padrão. O parâmetro α representa a diferença entre o tamanho médio dos domicílios e β a dispersão da proporção de domicílios por classe. Em geral, se:

$\alpha < 0$, o tamanho médio dos domicílios é menor que no padrão,

$\alpha > 0$, o tamanho médio dos domicílios é maior que no padrão,

$\beta < 1$, a concentração dos domicílios por classe é menor que no padrão,

$\beta > 1$, a concentração dos domicílios é maior que no padrão.

Em geral, espera-se que exista uma relação entre α e β. À medida que os domicílios menores são mais freqüentes (e portanto α é menor), a concentração dos domicílios por classe aumenta (e portanto β é maior). Tal concentração acontece, geralmente, nos domicílios de 2 a 3 pessoas (Kono, 1981; United Nations, 1973). Esta é uma tendência geral, que possibilita o estudo da coerência da informação de um país que experimenta mudanças no tamanho médio e na estrutura dos seus domicílios. Entretanto, a realidade é muito mais complexa e é possível que existam combinações diferentes das descritas.

As propriedades dos parâmetros α e β apresentados acima justificam o uso do modelo de desagregação por classes de tamanho, a partir de uma projeção disponível do número total de domicílios. O exercício de desagregação consistirá simplesmente em projetar tais parâmetros, a partir de dois métodos: o primeiro baseia-se na extrapolação de tendências passadas, que pressupõe a existência de uma série histórica de distribuição dos domicílios por classes de tamanho. A distribuição mais recente é utilizada como padrão ($\alpha = 0$ e $\beta = 1$) e são calculados os parâmetros para a série de distribuição passadas. A seguir, as tendências observadas com respeito aos parâmetros são extrapolados para cada ano ou qüinqüênio do período da projeção. O segundo método faz uso de uma distribuição meta ou limite, adotada para o último ano do período da projeção. Como na alternativa anterior, a distribuição mais recente pode ser tomada como padrão. A seguir, calculam-se os parâmetros α e β para a distribuição limite e interpola-se entre eles e os correspondentes a distribuição padrão para cada ano ou qüinqüênio intermediário. A estrutura limite pode ser a observada

em algum outro país, ou região do mesmo país, para onde se supõe que a estrutura domiciliar observada possa tender.

Qualquer que seja o método utilizado, dispor-se-á de um parâmetro α e de um parâmetro β para cada ano ou qüinqüênio do período da projeção. Tais parâmetros, mais os valores y_x^p, permitem obter a distribuição dos valores y_x para cada ano ou qüinqüênio pela fórmula (6). Neste caso, entretanto, os parâmetros são função do tempo:

$$y_{x,t+n} = \alpha_{t+n} + \beta_{t+n} y_{x,t}^p \tag{2.7}$$

Onde α_{t+n} e β_{t+n} são estimados para um momento dado da projeção (ano $t+n$, sendo t o ano inicial) e $y_{x,t}^p$ são os duplos logaritmos das projeções acumuladas de domicílios por classes de tamanho correspondentes à distribuição padrão ($\log(-\log D_x^p)$).

Os valores logaritmos das distribuições obtidos são transformados em proporções acumuladas pela equação:

$$D_{x,t+n} = e^{-e^{y_{x,t+n}}} \tag{2.8}$$

Finalmente, as proporções acumuladas transformam-se em proporções simples, as quais são aplicadas ao número total de domicílios previamente projetado, obtendo-se, assim, o número absoluto de domicílios por classes de tamanho. Este exercício repete-se para cada ano ou qüinqüênio do período da projeção.

A principal vantagem deste método é que os parâmetros podem ser projetados com flexibilidade, propondo uma evolução futura da distribuição dos domicílios por classes de tamanho coerente com tendências passadas, com as experiências de outros países ou com qualquer hipótese considerada razoável.

O desenvolvimento aqui apresentado é similar à aplicação da função de Gompertz na representação das taxas específicas de fecundidade acumuladas (CELADE, 1984) e ao que Brass tem desenvolvido na área da mortalidade através do sistema logito (United Nations, 1983).

3 – AVALIAÇÃO DO MODELO

Uma forma de avaliar um modelo é comparando séries de valores observados com séries de valores gerados pelo modelo. As diferenças entre ambas as séries indicam sua eficiência ou adequabilidade.

Numa primeira avaliação, o modelo foi aplicado às distribuições relativas dos domicílios por classes de tamanho correspondentes às principais áreas do mundo (ver Tabela 1). Cabe assinalar que existe uma considerável variação entre as áreas, polarizadas entre as desenvolvidas, onde prevalecem domicílios pequenos, e as subdesenvolvidas, onde domicílios maiores são mais freqüentes (ver Gráfico 1). Estas variações

não refletem apenas diferenças nos níveis de fecundidade, ainda que este seja provavelmente o fator mais importante, mas também diferenças nas estruturas etárias das respectivas populações. Em geral, os países subdesenvolvidos apresentam alta concentração populacional abaixo dos 20 anos. Já que a probabilidade de formar e manter um domicílio nessa faixa etária é reduzir, o tamanho dos domicílios tende a ser maior que nos países desenvolvidos, onde a proporção da população abaixo de 20 anos é sensivelmente menor. Uma outra dimensão a ser considerada é o grau de "nuclearização" das famílias, que nos países desenvolvidos é maior que nos subdesenvolvidos.

As percentagens da Tabela 1 foram acumuladas e determinaram-se os parâmetros α e β para as distribuições correspondentes a cada área, utilizando como padrão a dos países desenvolvidos.

Para isto, calcularam-se os duplos logaritmos das freqüências relativas acumuladas (y_x e y_x^p) e os αs e os βs respectivos pelo método dos mínimos quadrados. Os resultados deste exercício estão apresentados na Tabela 2.

A fim de avaliar o modelo de forma sistemática, utilizou-se o coeficiente de Theil (1971), que quantifica as diferenças entre séries de dados observados e estimados[2]. Com base nas informações da Tabela 2 (αs, βs e a série de valores y_x^p), e nas equações (7) e (8), foram estimadas as distribuições percentuais dos domicílios por classes de tamanho para cada área. Os resultados destas estimativas são apresentados na Tabela 3, junto com as respectivas percentagens observadas. Na mesma tabela são apresentados ainda os resultados da aplicação do coeficiente de Theil. Para todas as áreas consideradas, os valores são baixos, o que sugere que o modelo utilizado para as estimativas é adequado. Também é mostrada a decomposição do coeficiente em três componentes, cada um associado a um tipo diferente de erro. O erro de tendência refere-se às diferenças entre as médias das séries estimada e observada e, portanto, está relacionado a divergências na direção global destas. O erro de desvio diz respeito às diferenças entre os desvios padrão e refere-se, em um sentido amplo, à divergência nas formas das distribuições das duas séries. Por fim, o erro de correlação resume as diferenças entre os valores individuais das duas séries. Em todos os casos, os erros de tendência são nulos e os de desvio e de correlação extremamente baixos. Esses resultados confirmam que o modelo é um instrumento adequado de representação de uma distribuição relativa de domicílios por classes de tamanho.

Um exercício similar ao anterior foi realizado, desta feita aplicando-se o modelo às distribuições dos domicílios por classes de tamanho, segundo os Censos brasileiros de 1960, 1970 e 1980 (ver Tabela 4 e Gráfico 2).

Entre 1960 e 1970, as distribuições não apresentam maiores diferenças. Já em 1980, aumentou a percentagem dos domicílios de 2 e 4 membros e diminuiu a de 7 e mais. Houve incremento, também, na percentagem dos domicílios unipessoais. Em 1960 e 1970, o número médio de pessoas por domicílio chegaria a 5,19 e 5,28, respectivamente; em 1980 alcançou 4,70. O principal determinante dessas tendências tem sido, certamente, a substancial queda da fecundidade experimentada pela população brasileira desde a segunda metade da década de 60 (Arretx, 1984; Carvalho, 1985; Fernandez e

Carvalho, 1986; Merrik, 1985).

As percentagens da Tabela 4 foram acumuladas e, com base na mesma metodologia do exercício anterior, foram calculados os parâmetros respectivos para as distribuições de cada ano censitário, utilizando-se como padrão o ano de 1980 (ver Tabela 5). O modelo foi novamente avaliado, utilizando-se o coeficiente de Theil. A partir das informações da Tabela 5, foram estimadas as distribuições percentuais dos domicílios, por classes de tamanho para 1960 e 1970. Aplicando-se estas percentagens ao número total dos domicílios nestes anos, estimaram-se os números absolutos de domicílios por classes de tamanho (ver Tabela 6). Registrou-se novamente similaridade entre os valores estimados e os observados, ficando mais uma vez estabelecida a adequabilidade do modelo na apresentação de uma distribuição dos domicílios por classes de tamanho.

Esta adequabilidade da função de Gompertz para representar distribuições de domicílios por tamanho é fundamental na projeção de tais distribuições. Os métodos de projeções mais utilizados partem, normalmente, de séries estatísticas já observadas para determinar uma estrutura funcional de previsão. Como este método pode resumir distribuições passadas e presentes com apenas dois valores (parâmetros α e β), a evolução futura das ditas distribuições pode ser projetada com bastante flexibilidade.

4 – PROJEÇÃO DOS DOMICÍLIOS POR CLASSES DE TAMANHO PARA O BRASIL: 1980-2025

Uma projeção de domicílios é essencialmente uma estimativa das necessidades futuras de habitação resultantes de alterações no crescimento populacional, na estrutura etária da população, e na tendência dos adultos para formar seus próprios domicílios (vide Burch, 1980; Kuznetz, 1978). Como já dito anteriormente, as técnicas baseadas nas taxas específicas de chefes de domicílios são as mais recomendáveis para a projeção do número total de domicílios. Porém, estas técnicas não têm a mesma eficácia no caso específico dos dados censitários brasileiros. O motivo principal é que o conceito de chefe de domicílio varia entre os diferentes Censos, impossibilitando estabelecer as tendências passadas necessárias à projeção do número total de domicílios (Altmann, 1984). Assim, as projeções de domicílios realizadas no Brasil têm se baseado em outras técnicas, especialmente matemáticas.

Para este trabalho, utilizou-se o método desenvolvido por Frias (1987). Ele definiu que o número de domicílios seria estimado através de uma função que o relacionasse com a população da seguinte forma:

$$T_t = A + B\,P_t \tag{4.1}$$

Onde T_t é o número total de domicílios no ano t, P_t é a população total no ano t e A e B são parâmetros a serem determinados.

Com base nas informações censitárias sobre população e domicílios divulgadas em 1940, 1950, 1960, 1970 e 1980, Frias calculou os valores dos parâmetros A e B por regressão de mínimos quadrados, onde se verificam elevados valores da fração de variação explicada. O valor obtido para A foi -15.140.921,2 e para B foi 0,33908103. É importante assinalar que o parâmetro B representa o inverso do número médio de pessoas por domicílio. Esta característica do referido parâmetro permitiu controlar os valores limites da densidade domiciliar. O limite da relação população/domicílio $(1/B)$ foi 2,95.

Utilizando os parâmetros A e B e uma projeção populacional recente (Neupert, 1987), projetou-se o número total de domicílios para cada qüinqüênio entre 1980 e 2025, e também o número de pessoas por domicílio (ver Tabela 7).

Com base nas projeções anteriores do número de domicílios, este contingente foi projetado por classes de tamanho aplicando-se a função de Gompertz previamente descrita. Como distribuição padrão, utilizou-se a correspondente ao Censo de 1980, e como distribuição limite, a observada nos países desenvolvidos e que foi apresentada na Tabela 1. Supõe-se que a distribuição dos domicílios por classes de tamanho no Brasil deve tender para tal distribuição. Esta estrutura será alcançada aproximadamente no ano 2050.

As freqüências relativas acumuladas observadas nos países desenvolvidos e no Brasil em 1980, foram transformadas em duplos logaritmos (valores y_x e y_x^p) e, pelo método das médias, estimaram-se os parâmetros α e β para os anos 1980 e 2050. Os valores destes parâmetros para os qüinqüênios intermediários entre 1980 e 2025 foram obtidos por interpolação linear (ver Tabela 8).

A determinação estrutura limite no ano 2050, e não em 2025, que é o último ano do período da projeção, não foi aleatória. Segundo as projeções populacionais, aproximadamente no ano 2025 a população brasileira deverá alcançar um nível de reposição, ou seja, as taxas de fecundidade total atingirão um patamar um pouco superior ao de dois filhos por mulher. Entretanto, a experiência de países desenvolvidos, onde tal nível já foi alcançado, indica que, após ser atingido o nível de reposição, são necessários ainda entre 20 e 25 anos para se chegar a uma composição etária favorável à existência de proporções maiores de domicílios pequenos (Kono, 1981). A variação no número de domicílios está defasada em aproximadamente duas décadas com respeito a variações na fecundidade. Por outro lado, tendo em conta as tendências recentes, parece difícil que o Brasil chegue a ter, no final do período considerado, o mesmo grau de concentração de domicílios pequenos, especialmente unipessoais, que o observado nos países desenvolvidos. A estrutura domiciliar nestes últimos não reflete apenas um regime de baixa fecundidade e uma população com uma elevada proporção de adultos, características que a população brasileira certamente vai apresentar no futuro. Reflete, também, uma elevada propensão das pessoas que não estão em união conjugal a formar ou manter seus próprios domicílios. No entanto, segundo a Tabela 4, a percentagem de domicílios unipessoais no Brasil foi apenas 6,2%, cifra inferior à apresentada por qualquer das áreas subdesenvolvidas da Tabela 1. Cabe também assinalar que,

entre 1970 e 1980, este valor cresceu em apenas 1 ponto percentual. Estes dados indicariam que são principalmente as pessoas casadas as que formam novos domicílios, e que aquelas que não estão em união conjugal tendem a se agregar a outros domicílios. Sugerem, também, que seria difícil que no seja final do período da projeção, o Brasil chegue a ter, por exemplo, 16% de domicílios unipessoais. Tal cifra parece ser mais provável de ser atingida no ano 2050.

Voltando ao exercício, os valores projetados de α e β foram aplicados sucessivamente aos duplos logaritmos da distribuição padrão ($y^p_{x,1980}$ para $x = 1, 2, \cdots, 6$) segundo a equação (7) e obtiveram-se as distribuições projetadas em valores logarítmicos. Estes foram transformados em freqüências relativas acumuladas pela fórmula (8) e, em seguida, em freqüências relativas simples (ver Tabelas 9, 10 e 11).

É importante repetir que a população brasileira experimentou uma substancial queda da fecundidade durante as últimas décadas. A taxa de fecundidade total desceu de aproximadamente 6,2 filhos por mulher no começo dos anos 60 para 3,4 filhos no início da presente década (Fernandez e Carvalho, 1986; Oliveira e Silva, 1986; Wong, 1986). Esta queda resultou numa substancial desaceleração das taxas de crescimento populacional, que deverá estender-se às próximas décadas. Porém, quando este crescimento populacional é comparado com o do número total de domicílios (Tabela 11), é interessante notar que o primeiro é sempre menor que o último. Isto é, o resultado do fato de os indivíduos que formam novos domicílios num dado momento serem aqueles que nasceram aproximadamente duas décadas antes. Desse modo, o crescimento do número de domicílios está defasado em 20 a 25 anos com respeito ao crescimento populacional. Assim, apenas para a primeira década do próximo século, o crescimento do número total de domicílios será similar ao crescimento populacional previsto para a presente década. Porém, já é esperada uma desaceleração no ritmo de crescimento do número de domicílios para a presente e a próxima décadas, como resultado da queda da fecundidade ocorrida durante a segunda metade dos anos 60 e 70.

Com respeito aos domicílios por classes de tamanho, todas as classes vão experimentar crescimento durante o período da projeção. As maiores taxas anuais de crescimento incidem naqueles de uma só pessoa, seguidos pelos de duas e três pessoas. Os domicílios de cinco, seis e mais pessoas terão crescimento inferior ao correspondente ao número total de domicílios. Em termos relativos (Tabela 9), os domicílios de uma e duas pessoas, que em 1980 representavam aproximadamente 20% do total, em 2000 representarão 24% e no ano 2025 quase 30%. Domicílios de três ou com quatro pessoas não deverão sofrer mudanças substanciais com respeito à sua incidência, estabilizando-se em aproximadamente 35% do total durante todo o período da projeção. Os de cinco pessoas e mais, que em 1980 atingiam 45%, no ano 2000 alcançarão 41% e no ano 2025 aproximadamente 35%.

Ainda que o Brasil esteja experimentando uma desaceleração no crescimento populacional, cuja influência já se nota no ritmo de formação de novos domicílios, os dados acima apresentados indicam que as taxas de crescimento do número de domicílios são

ainda elevadas, ao menos quando comparadas às taxas de crescimento populacional. Elas serão inferiores a 2% apenas no começo do próximo século. Não é possível esperar que a recente queda da fecundidade resulte em alívio imediato das demandas habitacionais. Também é necessário lembrar que uma elevada proporção das atuais necessidades de moradia não tem sido atendida. Considerando que o ritmo de formação de novos domicílios permanecerá em patamares relativamente altos durante as próximas duas ou três décadas, não se prevê redução substancial na demanda para os próximos anos. Ainda vale a pena mencionar que as mudanças esperadas no tamanho dos domicílios vai implicar elevada demanda por unidades residenciais pequenas, fato este que sugere uma revisão no planejamento físico de novas moradias em termos de tamanho.

Finalmente, estas projeções consideraram apenas a dimensão demográfica da futura demanda habitacional e pressupõem que a tendência passada da relação população/domicílio será mantida no futuro e que as demandas por classes de tamanho inclinam-se a uma estrutura similar à existente nos países desenvolvidos. Numa economia capitalista, entretanto, tal demanda não depende apenas de aspectos populacionais, mas também da magnitude e da distribuição dos recursos econômicos das famílias e dos custos das unidades habitacionais.

5 – CONCLUSÕES

Este trabalho procurou demonstrar que a função de Gompertz é um modelo consistente para a desagregação de uma projeção do número total de domicílios por classes de tamanho. Sua principal vantagem é que utiliza dados geralmente disponíveis na maioria das publicações censitárias, tornando-o útil à análise e à projeção de alguns aspectos da composição domiciliar nos países subdesenvolvidos. Porém, este estudo deve ser considerado apenas um esforço inicial da utilização deste modelo para tal propósito. Outras pesquisas deveriam ser realizadas a fim de se obterem formas mais refinadas para sua aplicação neste campo. Seria também recomendável que se testasse esta técnica na análise do tamanho domiciliar em áreas pequenas. Outra sugestão não menos importante diz respeito à aplicação do modelo na descrição e projeção de domicílios por classes de tamanho apenas em função do número de adultos. Com este exercício, seria possível obterem-se informações relacionadas com a prevalência de domicílios com famílias nucleares, troncais ou estendidas. Em outras palavras, a fecundidade, como componente do tamanho domiciliar, seria controlada, e apenas a tendência dos adultos a formar seus próprios domicílios seria considerada.

Finalmente, cabe destacar que o método apresentado seria útil para estimativas mais completas das necessidades habitacionais futuras segundo o tamanho das moradias. Tais estimativas deveriam estar baseadas em modelos que levem em consideração não apenas as futuras demandas habitacionais resultantes de fatores demográficos, mas também algumas variáveis econômicas como a magnitude e a distribuição da renda das famílias, custos de unidades residenciais e investimentos públicos em moradias para a população de baixa renda.

GRÁFICO 1

DISTRIBUIÇÃO PERCENTUAL DOS DOMICÍLIOS POR CLASSES DE TAMANHO, SEGUNDO AS PRINCIPAIS ÁREAS DO MUNDO

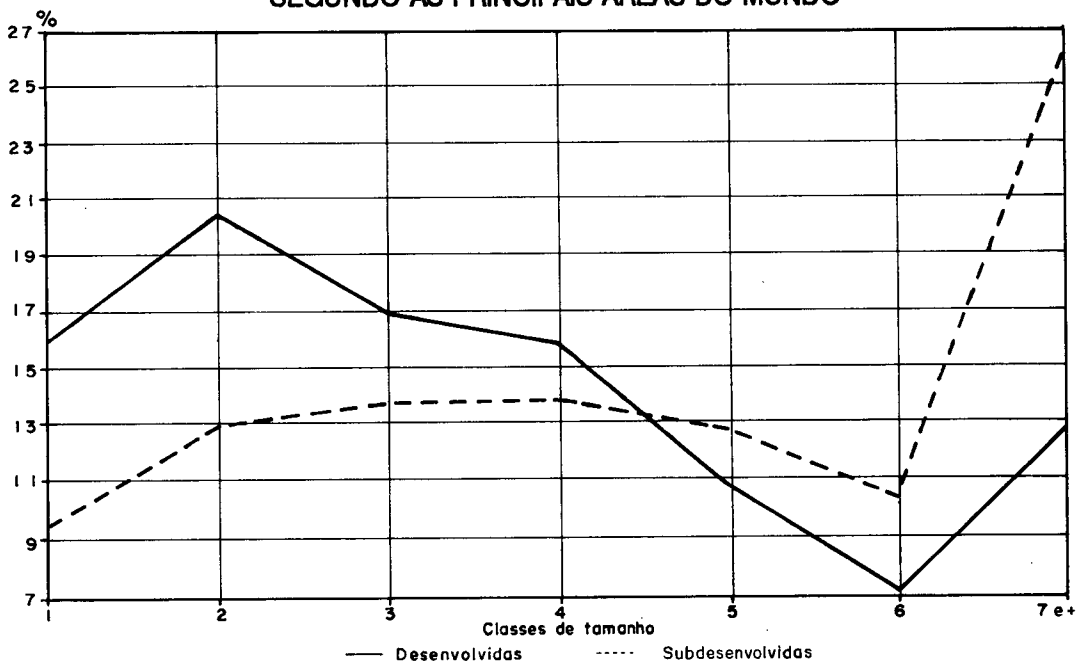


GRÁFICO 2

BRASIL: DISTRIBUIÇÃO PERCENTUAL DOS DOMICÍLIOS POR CLASSES DE TAMANHO, 1960, 1970 e 1980

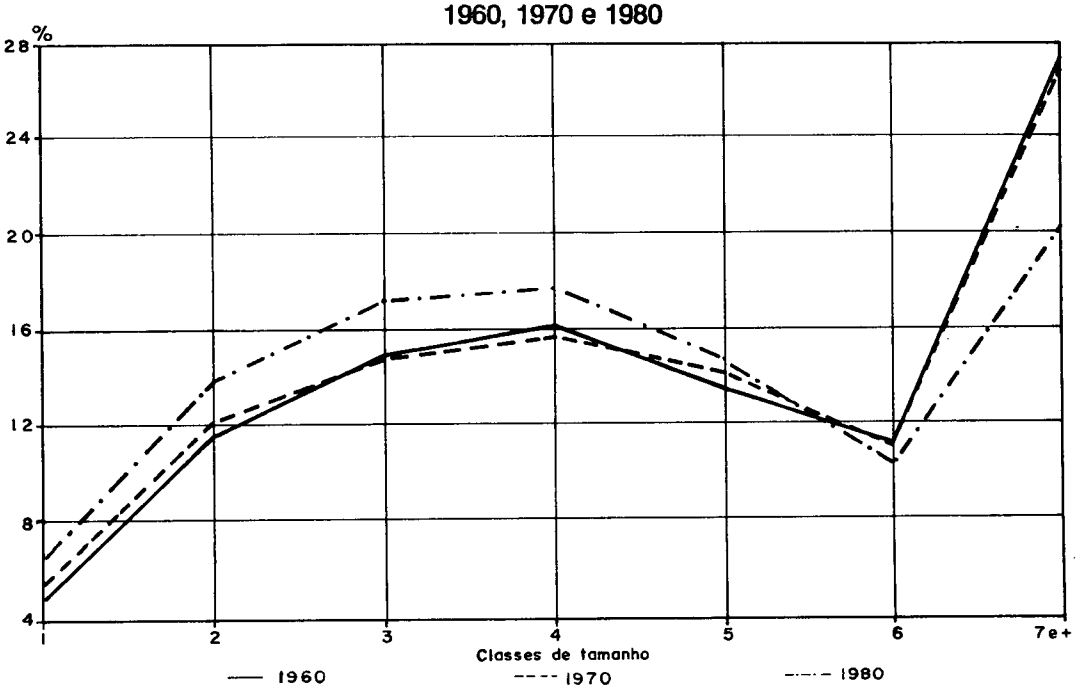


TABELA 1

DISTRIBUIÇÃO PERCENTUAL DOS DOMICÍLIOS POR CLASSES DE TAMANHO, SEGUNDO AS PRINCIPAIS ÁREAS DO MUNDO

ÁREAS	CLASSES DE TAMANHO							
	1	2	3	4	5	6	7 e mais	Total
África	11,2	16,9	15,7	13,7	11,7	9,3	21,5	100,00
América Latina	13,7	14,9	13,9	13,4	11,5	9,4	23,2	100,00
América do Norte	15,5	27,5	17,2	16,6	16,8	6,1	6,3	100,00
Ásia Oriental	11,6	12,1	14,6	15,6	12,9	11,1	22,1	100,00
Sul da Ásia	6,3	20,5	12,1	13,9	13,8	12,3	31,4	100,00
Europa	16,5	23,9	19,5	17,6	10,3	5,8	6,4	100,00
Oceania	11,3	14,9	14,1	15,3	12,9	8,5	23,0	100,00
Desenvolvidas	15,9	20,5	17,1	15,8	10,7	7,2	12,8	100,00
Subdesenvolvidas	9,4	13,1	13,8	13,9	12,8	10,6	26,4	100,00

FONTE - Kono (1981), Table 3.

TABELA 2

DUPLOS LOGARITMOS E VALORES DOS PARÂMETROS ALFA E BETA, SEGUNDO AS PRINCIPAIS ÁREAS DO MUNDO

ÁREAS	CLASSES DE TAMANHO							
	1	2	3	4	5	6	ALFA	BETA
África	0,7836	0,2385	-0,1917	-0,5917	-0,9992	-1,4185	0,2354	0,8273
América Latina	0,6870	0,2246	-0,1559	-0,5420	-0,9301	-1,3319	0,2135	0,7670
América do Norte	0,6229	-0,1696	-0,6783	-1,3319	-2,0220	-2,7323	-0,2834	1,1659
Ásia Ocidental	0,7674	0,3644	-0,0411	-0,4812	-0,9077	-1,3873	0,3219	0,8319
Sul da Ásia	1,0169	0,5888	0,2246	-0,1559	-0,5543	-0,9750	0,5721	0,7576
Europa	0,5888	-0,0983	-0,6685	-1,3669	-2,0394	-2,7160	-0,1273	1,3574
Oceania	0,7795	0,2922	-0,0956	-0,5328	-0,9720	-1,3418	0,2841	0,8237
Desenvolvidas	0,6091	0,0105	-0,4892	-1,0031	-1,4999	-1,9880	0,0000	1,0000
Subdesenvolvidas	0,8606	0,3999	0,0133	-0,3723	-0,7721	-1,1825	0,3857	0,7758

TABELA 3

COMPARAÇÃO ENTRE AS DISTRIBUIÇÕES DOS DOMICÍLIOS POR CLASSES DE TAMANHO OBSERVADAS NAS PRINCIPAIS ÁREAS DO MUNDO E AS ESTIMADAS PELA FUNÇÃO DE GOMPERTZ E CÁLCULO DO COEFICIENTE DE THEIL

ÁREAS	CLASSES DE TAMANHO								Coeficiente de Theil	Erro de Tendência	Erro de Desvio	Erro de Correlação
	1	2	3	4	5	6	7 e mais	Total				
África												
Observada	11,2	16,9	15,7	13,7	11,7	9,3	21,5	100,00				
Estimada	12,3	15,6	14,5	15,2	11,8	8,9	21,7	100,00	0,0334	0,0000	0,0036	0,0298
O-E	-1,1	1,3	1,2	-1,5	-0,1	0,4	-0,2	0,00				
América Latina												
Observada	13,7	14,9	13,9	13,4	11,5	9,4	23,2	100,00				
Estimada	13,9	14,8	13,5	14,2	11,2	8,8	23,6	100,00	0,0154	0,0000	0,0085	0,0068
O-E	-0,2	0,1	0,4	-0,8	0,3	0,6	-0,4	0,00				
América do Norte												
Observada	15,5	27,5	17,2	16,6	10,8	6,1	6,3	100,00				
Estimada	15,3	26,1	20,4	16,4	9,5	5,5	6,8	100,00	0,0455	0,0000	0,0019	0,0436
O-E	0,2	1,4	-3,2	0,2	1,3	0,6	-0,5	0,00				
Ásia Leste												
Observada	11,6	12,1	14,6	15,6	12,9	11,1	22,1	100,00				
Estimada	10,1	14,8	14,4	15,7	12,3	9,6	23,1	100,00	0,0446	0,0000	0,0228	0,0218
O-E	1,5	-2,7	0,2	-0,1	0,6	1,5	-1,0	0,00				
Sul da Ásia												
Observada	6,3	10,2	12,1	13,9	13,8	12,3	31,4	100,00				
Estimada	6,0	10,8	12,1	14,8	-12,9	10,9	32,5	100,00	0,0267	0,0000	0,0140	0,0127
O-E	0,3	-0,6	0,0	-0,9	0,9	1,4	-1,1	0,00				
Europa												
Observada	16,5	23,9	19,5	17,6	10,3	5,8	6,4	100,00				
Estimada	14,5	26,3	20,8	16,7	9,6	5,4	6,7	100,00	0,0430	0,0000	0,0219	0,0211
O-E	2,0	-2,4	-1,3	0,9	0,7	0,4	-0,3	0,00				
Oceania												
Observada	11,3	14,9	14,1	15,3	12,9	8,5	23,0	100,00				
Estimada	11,1	15,1	14,3	15,4	12,1	9,2	22,8	100,00	0,0145	0,0000	0,0038	0,0107
O-E	0,2	-0,2	-0,2	-0,1	0,8	-0,7	0,2	0,00				
Desenvolvidas												
Observada	15,9	20,5	17,1	15,8	10,7	7,2	12,8	100,00				
Estimada	15,9	20,5	17,1	15,8	10,7	7,2	12,8	100,00	0,0000	0,0000	0,0000	0,0000
O-E	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,00				
Subdesenvolvidas												
Observada	9,4	13,1	13,8	13,9	12,8	10,6	26,4	100,00				
Estimada	9,5	13,2	13,3	14,9	12,3	9,8	27,0	100,00	0,0199	0,0000	0,0100	,0098023
O-E	-0,1	-0,1	0,5	-1,0	0,5	0,8	-0,6	0,00				

TABELA 4

DISTRIBUIÇÃO DOS DOMICÍLIOS POR CLASSES DE TAMANHO - 1960-1980

Brasil

CLASSES DE TAMANHO	NÚMERO DE DOMICÍLIOS			PERCENTAGEM DE DOMICÍLIOS		
	1960	1970	1980	1960	1970	1980
1	645 873	916 119	1 566 061	4,78	5,19	6,19
2	1 580 457	2 129 348	3 449 677	11,70	12,07	13,64
3	2 019 934	2 622 422	4 316 185	14,96	14,86	17,06
4	2 196 564	2 763 694	4 492 639	16,27	15,66	17,76
5	1 836 914	2 505 637	3 699 707	13,60	14,20	14,63
6	1 533 126	1 961 268	2 588 744	11,35	11,12	10,23
7 e mais	3 691 757	4 744 899	5 180 398	27,34	26,89	20,48
Total	13 504 625	17 643 387	25 293 411	100,00	100,00	100,00

FONTE - IBGE, Censos Demográficos de 1960, 1970 e 1980.

TABELA 5

DUPLOS LOGARITMOS E VALORES DOS PARÂMETROS ALFA E BETA, SEGUNDO A DISTRIBUIÇÃO OBSERVADA EM 1980

CLASSES DE TAMANHO	ANOS		
	1960	1970	1980
1	1,1121	1,0847	1,0233
2	0,7722	0,5635	0,4812
3	0,1459	0,1272	-0,0028
4	-0,3011	-,03031	-0,5038
5	-0,7149	-0,7374	-1,0024
6	-1,1414	-1,1605	-1,4728
Alfa	0,2089	0,1476	0,0000
Beta	0,9346	0,8874	1,0000

TABELA 6

COMPARAÇÃO ENTRE AS DISTRIBUIÇÕES DOS DOMICÍLIOS POR CLASSES DE TAMANHO, SEGUNDO OS CENSOS BRASILEIROS DE 1960, 1970 E 1980 E AS ESTIMADAS PELA FUNÇÃO DE GOMPERTZ E CÁLCULO DO COEFICIENTE DE THEIL

ANOS	CLASSES DE TAMANHO								Coeficiente de Theil	Erro de Tendência	Erro de Desvio	Erro de Correlação
	1	2	3	4	5	6	7 e mais	Total				
1960												
Observada ...	645 873	1 580 457	2 019 934	2 196 564	1 836 914	1 533 126	3 691 757	13 504 625				
Estimada	546 937	1 408 532	1 994 633	2 305 239	2 077 011	1 561 135	3 611 138	13 504 625	0,0305	0,0000	0,0038	0,0268
O-E	98 936	171 925	25 301	-108 675	-240 097	-28 009	80 619	0				
1970												
Observada ...	916 119	2 129 348	2 622 422	2 763 694	2 505 637	1 961 268	4 744 899	17 643 387				
Estimada	996 851	1 988 410	2 567 113	2 856 464	2 551 234	1 933 715	4 749 600	17 643 387	0,0140	0,0000	0,0003	0,0137
O-E	-80 732	140 938	55 309	-92,770	-45 597	27 553	-4 701	0				
1980												
Observada ...	1 566 061	3 449 677	4 316 185	4 492 639	3 699 707	2 588 744	5 180 398	25 293 411				
Estimada	1 566 061	3 449 677	4 316 185	4 492 639	3 699 707	2 588 744	5 180 398	25 293 411	0,0000	0,0000	0,0000	0,0000
O-E	0	0	0	0	0	0	0	0				

TABELA 7

PROJEÇÃO DO NÚMERO TOTAL DE DOMICÍLIOS E DO NÚMERO DE PESSOAS POR DOMICÍLIO - 1980-2025

Brasil

ANOS	TOTAL DE DOMICÍLIOS	PESSOAS POR DOMICÍLIO
1980	25 293 411	4,66
1985	30 459 445	4,42
1990	34 802 800	4,23
1995	39 096 663	4,09
2000	43 322 597	3,98
2005	47 354 184	3,89
2010	51 169 598	3,82
2015	54 782 320	3,76
2020	58 158 702	3,72
2025	61 238 087	3,68

FONTE - Vide texto.

TABELA 8

PROJEÇÃO DOS PARÂMETROS ALFA E BETA - 1980-2050

Brasil

ANOS	ALFA	BETA
1980	0,00000	1,00000
1985	-0,03343	1,00253
1990	-0,06686	1,00507
1995	-0,10029	1,00760
2000	-0,13373	1,01014
2005	-0,16718	1,01268
2010	-0,20062	1,01522
2015	-0,23407	1,01776
2020	-0,26750	1,02030
2025	-0,30096	1,02283
2050	-0,46819	1,03550

TABELA 9

DISTRIBUIÇÃO PERCENTUAL PROJETADA DOS DOMICÍLIOS POR CLASSES DE TAMANHO - 1980-2025

Brasil

CLASSES DE TAMANHO	DISTRIBUIÇÃO PERCENTUAL PROJETADA DOS DOMICÍLIOS									
	1980	1985	1990	1995	2000	2005	2010	2015	2020	2025
1	6,19	6,74	7,31	7,92	8,55	9,21	9,90	10,62	11,37	12,15
2	13,64	14,14	14,63	15,10	15,57	16,01	16,45	16,87	17,27	17,65
3	17,06	17,25	17,41	17,56	17,68	17,79	17,87	17,93	17,98	18,01
4	17,76	17,67	17,56	17,44	17,31	17,16	17,00	16,83	16,65	16,46
5	14,63	14,40	14,16	13,92	13,67	13,42	13,17	12,92	12,67	12,41
6	10,23	9,99	9,76	9,53	9,30	9,07	8,84	8,62	8,40	8,18
7 e mais	20,48	19,82	19,17	18,54	17,93	17,34	16,76	16,20	15,66	15,13
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00

TABELA 10

PROJEÇÃO DO NÚMERO ABSOLUTO DE DOMICÍLIOS POR CLASSES DE TAMANHO - 1980-2025

Brasil

CLASSES DE TAMANHO	PROJEÇÃO DO NÚMERO ABSOLUTO DE DOMICÍLIOS									
	1980	1985	1990	1995	2000	2005	2010	2015	2020	2025
1	1 566 061	2 051 705	2 544 355	3 094 604	3 703 683	4 362 404	5 067 945	5 820 592	6 614 586	7 440 638
2	3 449 677	4 306 323	5 090 129	5 904 304	6 743 288	7 583 659	8 416 774	9 239 864	10 042 033	10 807 694
3	4 316 185	5 253 002	6 059 712	6 863 995	7 659 919	8 422 197	9 144 004	9 825 130	10 457 240	11 027 311
4	4 492 639	5 381 732	6 112 075	6 818 952	7 497 965	8 126 306	8 700 044	9 221 490	9 685 423	10 082 275
5	3 699 707	4 384 958	4 927 308	5 440 546	5 922 299	6 355 918	6 739 999	7 077 861	7 367 013	7 601 611
6	2 588 744	3 043 979	3 396 375	3 724 442	4 027 150	4 293 880	4 524 496	4 721 956	4 885 301	5 011 329
7 e mais	5 180 398	6 037 746	6 672 847	7 249 820	7 768 293	8 209 820	8 576 337	8 875 425	9 107 105	9 267 228
Total	25 293 411	30 459 445	34 802 800	39 096 663	43 322 597	47 354 184	51 169 598	54 782 320	58 158 702	61 238 087

TABELA 11

TAXAS ANUAIS DE CRESCIMENTO DA POPULAÇÃO E DO NÚMERO DE DOMICÍLIOS POR CLASSES DE TAMANHO - 1980-2025

Brasil

CLASSES DE TAMANHO	TAXAS ANUAIS								
	1980-85	1985-90	1990-95	1995-2000	2000-2005	2005-2010	2010-2015	2015-2020	2020-2025
1	5,55	4,40	3,99	3,66	3,33	3,04	2,81	2,59	2,38
2	4,54	3,40	3,01	2,69	2,38	2,11	1,88	1,68	1,48
3	4,01	2,90	2,52	2,22	1,92	1,66	1,45	1,25	1,07
4	3,68	2,58	2,21	1,92	1,62	1,37	1,17	0,99	0,81
5	3,46	2,36	2,00	1,71	1,42	1,18	0,98	0,80	0,63
6	3,29	2,22	1,86	1,58	1,29	1,05	0,86	0,68	0,51
7 e mais	3,11	2,02	1,67	1,39	1,11	0,88	0,69	0,52	0,35
Total	3,79	2,70	2,35	2,07	1,80	1,56	1,37	1,20	1,04
População	1,96	1,82	1,65	1,50	1,33	1,19	1,06	0,94	0,82

NOTAS

1 - A série de valores y_x divide-se em dois grupos aproximadamente iguais. Para cada grupo calcula-se a média aritmética: y_1 e y_2. O mesmo exercício é feito para os valores y_x^p a fim de obter $\bar{y}_1^p$ e $\bar{y}_2^p$. Os dois parâmetros calculam-se com as seguintes equações:

$$\beta = (\bar{y}_2 - \bar{y}_1)/(\bar{y}_2^p - \bar{y}_1^p) \quad \text{e} \quad \alpha = y_1 - \beta\,\bar{y}_1^p$$

2 - A fórmula do coeficiente de Theil é:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \Big/ \left(\sqrt{\sum_{i=1}^{n} x_i^2} + \sqrt{\sum_{i=1}^{n} \hat{x}_i^2}\right)$$

onde $\hat{x}_i$ é a série de valores estimados e x_i é a série de valores observados.

Quando $x_i = \hat{x}_i, d = 0$

O coeficiente pode ser desagregado em três partes, cada uma indicando um tipo específico de erro:

Erro de tendência: $e_t^2 = \left((\bar{x} - \bar{\hat{x}})/D\right)^2$

Erro de desvio: $e_d^2 = \left((\sigma - \hat{\sigma})/D\right)^2$

Erro de correlação: $e_c^2 = \left(2^*(1-\rho)^*\sigma^*\hat{\sigma}\right)/D^2$

Onde:

$$D = \sqrt{1/n^* \sum_{i=1}^{n} x_i^2} + \sqrt{1/n \sum_{i=1}^{n} \hat{x}_i^2}$$

BIBLIOGRAFIA

ALTMAN A. M. G. A informação da família no censo demográfico. In: SEMINÁRIO METODOLÓGICO DOS CENSOS DEMOGRÁFICOS, 3, 1984, Outro Preto, MG. *Censos, Consensos, Contra-Sensos*, Ouro Preto, MG : Associação Brasileira de Estudos Populacionais, 1984, p. 235-281.

ARRETX, C. Proyección de la población de Brasil, por sexo y grupos quinquenales de edad, 1950-2150. In: METODOS para proyecciones demograficas. San José, Costa Rica : CELADE, 1984.

BONGAARTS, J. Simulation of the family life cycle. In: INTERNATIONAL POPULATION CONFERENCE, Manila, 1981. Liège : IUSSP, 1981, v. 3, p. 399-416.

BRASS, W. Perspective in population prediction: illustrated by the statistics of England and Walles. *Journal of the Royal Statistical Society, S.A*, v. 137, n. 4, p. 532-583, 1974.

BURCH, T. The index of overall headship: a simple measure of household complexity standardized for age and sex. *Demography*, v. 17, n. 1, p. 25-37, 1980.

CARVALHO, J. A. M. The new demographic profile of Brazil. *Brazilian Economic Studies*, Rio de Janeiro, n. 6, p. 237-261, 1985.

FARID, S. M. "On the pattern of cohort fertility", *Population Studies*, v. 27, n. 1, p. 159-168, 1973.

FERNANDEZ, R. E., CARVALHO, J. A. M. A evolução da fecundidade no Brasil: período 1957-1979. *Revista Brasileira de Estudos de População*, v. 3, n. 3, p. 67-86, 1986.

FRIAS, L. A. Projeções da população residente e do número de domicílios particulares ocupados por situação urbana e rural, segundo as unidades da federação no período 1985-2020. In: FUTURO da população Brasileira: projeções, previsões e técnicas. São Paulo : Associação Brasileira de Estudos Populacionais, 1987, p. 148-173.

GOMPERTZ, B. On the nature of the function expressive of the law of human mortality, and on the new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society*, p. 513-585, 1825.

KONO, S. Further contrivances on methods of households projections with special attention to household size and to social development planning. In: INTERNATIONAL POPULATION CONFERENCE, Manila, 1981. Liège : IUSSP, 1981, v. 3, p. 485-502.

KUZNETZ, S. Size and age structure of family households: exploratory comparisons. *Population and Development Review*, v. 4, p. 187-223, 1978.

MARTIN, M. P. D. Une application des fonctions de Gompertz à l'étude de la fecondite d'une cohorte. *Population*, v. 22, p. 1085-1096, 1967.

MENKEN, J. Age and fertility: How late can you wait? *Demography*, v. 22, n. 4, p. 469-483, 1985.

MERRICK, T. *Recent fertility declines in Brazil, Colombia, and Mexico.* Washington, D.C. : World Bank, 1985. (Wold Bank Staff Working Papers, n. 692).

METHODS of projecting households and families: Manual 7. New York : United Nations, 1973.

METODOS para proyecciones demográficas. San José, Costa Rica : CELADE, 1984.

NEUPERT, R. Nova projeção da população brasileira: hipóteses baseadas em informações recentes. In: FUTURO da população brasileira: projeções, previsões e técnicas. São Paulo : Associação Brasileira de Estudos Populacionais, 1987, p. 52-87.

OLIVEIRA, L. A. P., SILVA, N. L. P. Tendências da fecundidade nos primeiros anos da década de 80. In: ENCONTRO NACIONAL DE ESTUDOS POPULACIONAIS, 5, 1986, Águas de São Pedro, SP, *Anais...* Águas de São Pedro, SP : ABEP, 1986, p. 213-232.

ROMANIUK, A., TAMMY, S.M. *Projections of incomplete cohort fertility for Canada by means of the Gompertz function: analitical and technical memorandum*, n. 1. Ottawa. Canada : Census Division, 1969.

THEIL, H. *Principles of Econometrics.* New York : John Wiley, 1971.

WATKINS, S. C., MENKEN, J., BONGAARTS, J. *Continuities and canges in the American family.* California, 1984. Trabalho Apresentado no Encontro Anual da Social Science History Association.

WONG, L. R. A diminuição dos nascimentos e a queda da fecundidade no Brasil dos anos pós 80. In: ENCONTRO NACIONAL DE ESTUDOS POPULACIONAIS, 5, 1986, Águas de São Pedro, SP, *Anais...* Águas de São Pedro, SP : ABEP, 1986, v. 1, p. 233-256.

WUNSCH, G. Courbes de Gompertz et perspectives de fécondité. *Recherches Economiques*, v. 32, p. 457-468, 1966.

Yi, S. Changes in family structure in China. *Population and Development Review*, v. 12, n. 4, p. 675-703, 1986.

RESUMO

O objetivo do presente trabalho é propor uma técnica, baseada no uso da função de Gompertz, para representar o número de domicílios por classes de tamanho. Por extensão, também poderia ser utilizada para desagregar, por classes de tamanho, uma projeção disponível do número total de domicílios. Esta técnica não considera diretamente os aspectos dinâmicos do ciclo de vida familiar, porém apresenta uma vantagem metodológica: utiliza informações normalmente disponíveis na maioria das publicações censitárias.

O trabalho tem início com a apresentação das características gerais da função de Gompertz e da forma que a mesma pode ser utilizada na análise e projeção dos domicílios por classes de tamanho. A seguir, são avaliadas suas possibilidades analíticas com dados internacionais recentes e com dados dos Censos brasileiros de 1960, 1970 e 1980. Por

fim, a função é utilizada para desagregar, por classes de tamanho, uma projeção recente do número total de domicílios no Brasil até o ano 2025.

As respectivas análises mostram que, de fato, a função de Gompertz é um modelo consistente para desagregar, por classes de tamanho, o número total de domicílios. O método parece ser útil como base para o desenvolvimento de modelos para estimar as necessidades habitacionais futuras, segundo o tamanho das moradias.

ABSTRACT

The purpose of this paper is to present a technique, based on the Gompertz relational function, to represent a distribution of households by size class. By extension, it may also be used to disagregate, by size class, an available projection of the total number of households. The technique does not consider directly the dynamic aspects of the family life cycle, however, it has a methodological advantage: it makes use of information usually available in most census publications. The general characteristics of the Gompertz function, and the way in which it may be used in the analysis and projections of households by size class, are initially presented. Following, their analytical possibilities are evaluated with international data and with data from the 1960, 1970, and 1980 Brazilian censuses. Finally, the function is used to disagregate, by size class, a recent projection of households for Brazil. The analyses show that, in fact, the Gompertz function is a consistent model to disagregate the total number of households by size class. The methods may be usefull as a base to develop models to estimate future housing needs according to the size of the dwellings.

RESENHAS BIBLIOGRÁFICAS

– ROBUST REGRESSION AND OUTLIER DETECTION. Peter J. Rousseeuw and Annick M. Leroy. Wiley. New York. 1987. xiv + 329 pages.

A Análise de Regressão é, sem dúvida, uma das mais importantes ferramentas estatísticas, utilizada em forma rotineira na maioria das ciências aplicadas. Entre as muitas possíveis técnicas de regressão, as baseadas no método de mínimos quadrados tem recebido a maior atenção, um pouco pela facilidade de cálculo e "um muito" pela tradição.

Felizmente é cada vez maior a consciência entre os cientistas que uma aplicação cega dessa técnica pode induzir a conclusões totalmente erradas, em face da ocorrência, muito comum na prática, de "outliers" (observações muito afastadas da maioria) e/ou de "pontos influentes". Freqüentemente tais observações não são detectadas pelas clássicas "técnicas de diagnósticos", principalmente porque elas mesmas estão baseadas em ajustes por mínimos quadrados.

Procurando remediar essa situação, estão sendo desenvolvidas novas técnicas estatísticas que não são afetadas pela presença dessas "observações anômalas". Tais métodos têm recebido o nome de "robustos" ou "resistentes". As conclusões que deles se pode tirar são confiáveis, embora a amostra tenha uma certa proporção de "dados contaminados".

A principal mensagem deste livro dos professores Rousseeuw e Leroy é que as "técnicas de regressão robusta" são muito úteis para identificar e tratar essas "amostras

contaminadas". É por essa razão que a publicação deste trabalho deve ser celebrada por toda a comunidade estatística.

Os autores esperam que seu trabalho sirva para conciliar duas posições "filosóficas" aparentemente contrapostas entre os analistas de dados. Por um lado, estão os que preferem usar as "técnicas de diagnóstico de "outliers" baseadas em um ajuste preliminar por mínimos quadrados, identificando, dessa maneira,esses pontos "ruins", tirando-os da amostra e finalmente ajustando-os à amostra restante numa regressão pelo mesmo método de mínimos quadrados. Muitos entre esses estatísticos acham que dessa maneira evita-se certa "arbitrariedade" que pensam eles têm alguns métodos robustos, embora esse procedimento de identificar pontos "ruins"e rejeitá-los já seja uma técnica robusta. Por outro lado, estão os "robustos", isto é, aqueles que consideram arbitrário qualquer procedimento que descarte observações genuínas da amostra, e, daí, prefiram usar uma técnica que faça justiça a todos os dados, embora certos procedimentos robustos impliquem a rejeição de pontos muito afastados.

Rousseeuw e Leroy propõem usar ambos os pontos de vista num mesmo procedimento: fazer primeiramente um ajuste robusto, identificar como "outliers" aqueles pontos cujos resíduos com respeito a esse ajuste ultrapasse certo limite, e concluir com um ajuste de mínimos quadrados ponderados, colocando peso zero (rejeição) ou peso menor que 1 nesses pontos. O livro todo é uma análise bastante completa e muito clara desta técnica.

O propósito dos autores, explicitamente declarado, foi o de facilitar a aplicação de técnicas de regressão robusta no cotidiano do trabalho de um estatístico aplicado.

Com essa perspectiva em mente, o livro foi escrito de maneira a fazê-lo resultar compreensível e didático para alunos e profissionais da estatística. As partes matematicamente técnicas, porém, não foram excluídas e aparecem assinaladas ao longo do texto com um (*) e pode-se omitir seu estudo sem atrapalhar a compreensão da utilidade e a aplicação da metodologia exposta. O material é todo organizado para ser usado como livro-texto em um segundo curso de regressão ou análise de dados.

Embora nenhum conhecimento prévio seja estritamente necessário, é óbvio que o melhor proveito do estudo deste livro será obtido quando já se conhecerem as técnicas clássicas de regressão. O autor desta crítica tem usado parte do livro em dois cursos de Mestrado em Estatística no IMPA, com bons resultados.

No entanto, um aspecto que não chega a ser muito negativo é o da ênfase que os autores dão a uma particular técnica de ajuste robusto, baseada no chamado "estimador do mínimo das medianas dos quadrados dos resíduos" (Least Median Square estimator). Mas também são analisados outros métodos clássicos e robustos, e daí o livro todo continuar a ser recomendável.

Rousseeuw e Leroy expressam sua esperança de que as técnicas por eles estudadas, e outras da mesma natureza,sejam incorporadas futuramente aos "pacotes" computacionais de uso mais freqüente nos "laboratórios" de estatística. Entretanto, eles

oferecem o programa usado para resolver os exemplos e problemas de aplicação que aparecem no livro. Esse programa chama-se PROGRESS e pode ser obtido dos autores escrevendo-se para: ROUSSEEUW, Peter, J. - Seminar of Statistics, ISES - University of Fribourg - CH 1700 Fribourg - Suiça ou, também, para o autor desta crítica.

Para terminar, vejamos o conteúdo do texto, apenas um esquema com o cabeçalho de cada capítulo, seguido de um breve comentário.

CHAPTER I: INTRODUCTION

São introduzidos o modelo de regressão linear múltipla a ser estudado e certas definições referentes a "outliers", "técnicas de diagnóstico", "ponto de ruptura", "estimadores robustos", etc.

Uma boa e cuidada escolha de "exemplos de aplicação" esclarece a utilidade de tais conceitos na análise de dados.

CHAPTER II: SIMPLE REGRESSION

Incluído por razões didáticas, é possível inserir o conteúdo deste capítulo em cursos de estatística geral ainda num nível introdutório.

CHAPTER III: MULTIPLE REGRESSION

Constitui uma das partes essenciais do livro. Muitos exemplos e detalhada explicação do uso do programa PROGRESS fazem o estudo das três primeiras seções deste capítulo muito valioso para qualquer curso de regressão orientado para as aplicações.

CHAPTER IV: THE SPECIAL CASE OF ONE-DIMENSIONAL LOCATION

O objetivo deste capítulo é introduzir o leitor no estudo da matemática envolvida na prova das propriedades assintóticas dos estimadores robustos em geral e do já citado LMS-estimador, em particular. Sua leitura é totalmente prescindível num curso de interesse prático.

CHAPTER V: ALGORITHMS

Consiste em uma minuciosa explicação do algoritmo que implementa o cálculo do LMS-estimator, de interesse apenas em cursos orientados para a Estatística Computacional.

CHAPTER VI: OUTLIER DIAGNOSTICS

Junto com o Capítulo III, forma o "coração" do livro. É uma excelente e muito didática exposição das técnicas de diagnósticos de "outliers", tanto clássicas como "resistentes", que tem recebido a maior aceitação por parte dos estatísticos "aplicados".

CHAPTER VII: RELATED STATISTICAL TECHNIQUES

Traz um conjunto de valiosas sugestões visando à extensão das técnicas expostas a outros modelos: Análise Multivariada, Séries Temporais, etc.

Todos os capítulos terminam por uma boa lista de problemas, tanto de complementação teórica e prática como de aplicação do exposto no texto. Também aparecem propostos alguns problemas abertos.

Em resumo, é recomendável que este livro forme parte da biblioteca de todo es-

tatístico preocupado em manter-se atualizado sobre os avanços das pesquisas na Estatística em geral.

(Oscar H. Bustos - IMPA - Rio de Janeiro, RJ)

– PARAMETRIC STATISTICAL MODELS AND LIKELIHOOD. Ole E. Barndorff-Nielsen (Springer Lecture Notes in Statistics, 50, 1988, 276 p.)

A ciência Estatística deve a R. A. Fisher muitos de seus conceitos fundamentais tais como verossimilhança, informação, suficiência e ancilaridade. Mas, as idéias de Fisher eram, muitas vezes, vagas ou incompletas, e o desenvolvimento da teoria fisheriana continua até hoje. Um dos principais contribuidores no desenvolvimento desta teoria, nos últimos tempos, é o Professor Ole E. Barndorff-Nielsen, e o seu livro é essencialmente um resumo da pesquisa que vem realizando ao longo desta década. O livro retrata os últimos avanços na inferência teórica fisheriana, incluindo aspectos de Geometria Diferencial.

Um conceito-chave da teoria fisheriana é o princípio de ancilaridade, segundo o qual a inferência deve ser feita na distribuição condicional, dada uma estatística ancilar. Uma estatística é dita ancilar quando sua distribuição não depende dos parâmetros do modelo. Este princípio está intimamente relacionado com o princípio de suficiência pelo qual a inferência deve ser feita utilizando-se a distribuição marginal de uma estatística suficiente.

Enquanto que o princípio de suficiência é quase unanimemente aceito, e também tem uso em outras teorias estatísticas, o mesmo não acontece com o princípio de ancilaridade. Este princípio pode ser considerado polêmico, e é muitas vezes, sujeito a discussões no âmbito da Estatística Teórica. O principal argumento contra ele é o de que o procedimento condicional é menos eficiente, em média, em repetições hipotéticas, quando comparado com o procedimento não-condicional. Os argumentos a favor do princípio de ancilaridade são: (i) que uma estatística ancilar é uma fonte de ruído; (ii) que o procedimento condicional não só elimina este ruído, mas também contribui para determinar o verdadeiro conteúdo de informação na amostra sobre os parâmetros; e (iii) que repetições hipotéticas do experimento são irrelevantes para a interpretação da amostra em questão.

Se aceitarmos o princípio de ancilaridade, ainda assim nos restam dois problemas a resolver. O primeiro é que nem sempre existe uma estatística ancilar e, mesmo quando existe, pode ser difícil trabalhar com a distribuição condicional correspondente do ponto de vista prático ou numérico. Uma das principais contribuições de Barndorff-Nielsen, e um dos pontos principais do livro, é uma fórmula chamada de p^* (fórmula de

Barndorff-Nielsen), que de uma só vez resolve ambos os problemas acima mencionados. Esta fórmula foi introduzida por Barndorff-Nielsen em um artigo publicado em 1983 na revista Biometrika.

A primeira contribuição de Barndorff-Nielsen em relação ao princípio de ancilaridade é relativamente simples: caso uma estatística ancilar não exista, basta procurar uma estatística que seja aproximadamente ancilar. Tal estatística pode ser encontrada facilmente através do vetor escore padronizado, que tem média aproximadamente igual a zero e matriz de variâncias e covariâncias aproximadamente igual à matriz identidade e, portanto, uma distribuição aproximadamente independente dos parâmetros do modelo.

A segunda e principal contribuição de Barndorff-Nielsen, a fórmula p^*, resolve a questão do cálculo aproximado da distribuição condicional de $\hat{\varphi}$, dado a, onde $\hat{\varphi}$ é o estimador de máxima verossimilhança do parâmetro φ, e a é a estatística ancilar.

A fórmula é bastante simples, utilizando somente a função de verossimilhança e a informação observada, ou seja, a segunda derivada da log-verossimilhança. Por incrível que pareça, é totalmente geral, e mesmo assim dá a solução exata para um grande número de exemplos, e em muitos outros ela fornece uma aproximação muito boa.

A fórmula é uma generalização da aproximação normal para distribuição de $\hat{\varphi}$, mas possui várias características que a aproximação normal não tem, sendo elas: (i) invariância sobre reparametrização; (ii) um erro de ordem $n^{-\frac{3}{2}}$; e (iii) convergência em geral uniforme sobre subconjuntos compactos. Em outras palavras, a aproximação obtida através da fórmula p^* é pelo menos "uma ordem de magnitude" melhor do que a da aproximação normal.

No livro, depois de um primeiro capítulo introdutório, são apresentados, no Capítulo 2, dois tipos de modelos, que são essenciais para se entender a fórmula p^* . O primeiro tipo de modelo é a classe de famílias exponenciais, onde a fórmula p^* reduz-se à chamada aproximação de ponto de sela, a qual em geral é muito boa, e, além disto, possui a propriedade citada de convergir uniformemente em subconjuntos compactos.

O segundo tipo de modelo introduzido no Capítulo 2 é a classe modelos gerados por um grupo de transformações do espaço amostral. Barndorff-Nielsen mostra que a fórmula é exata para todos os modelos essencialmente transformacionais, tornando esta classe de modelos o segundo exemplo onde a fórmula pode ser aplicada com sucesso. Um exemplo deste tipo de modelo é o de locação e escala, onde a fórmula de Barndorff-Nielsen é equivalente a uma fórmula de Fisher para a distribuição condicional de $\hat{\varphi}$, dada a configuração da amostra.

Além dos dois tipos de modelos acima mencionados, Barndorff-Nielsen mostra muitos outros exemplos onde a fórmula ou é exata ou fornece uma aproximação muito boa. Para mencionar um único exemplo, no caso da distribuição gama, a aproximação de Barndorff-Nielsen reduz-se à famosa fórmula de Stirling, que é conhecida como uma aproximação excelente para a função gama, para quase todos os reais positivos. Estes

exemplos e vários outros aspectos da fórmula p^* são tratados nos Capítulos 2, 6 e 7 do livro.

Os Capítulos 3 e 4 tratam de um assunto que está ganhando cada vez mais importância na estatística teórica, a Geometria Diferencial. Segundo o enfoque baseado na Geometria Diferencial, um modelo estatístico é uma variedade diferenciável, munida de várias estruturas geométricas, entre elas a geometria da matriz de Informação de Fisher. Este enfoque é bastante útil nos cálculos associados à fórmula p^* e outras expansões relacionadas com a função de verossimilhança. As diversas estruturas geométricas (tensores, conexões, cordas, etc.) estão ligadas à estrutura de momentos e cumulantes de derivadas da função de log-verossimilhança. Em particular, o Capítulo 5 trata de cumulantes e momentos em geral.

Um dos frutos da aplicação da Geometria Diferencial à Estatística é o de que ela facilita o cálculo de correções para testes de razão de verossimilhança ou testes de escore. Em particular, foi mostrado por Barndorff-Nielsen e Cox que existe uma relação íntima entre a fórmula p^* e a correção de Bartlett para o teste de razão de verossimilhança. A fórmula p^* pode ser utilizada para derivar vários outros tipos de procedimentos aproximativos para inferência. Um exemplo disto é a noção de verossimilhança perfilada modificada, que, dentre outras coisas, é tratada no Capítulo 7.

O livro tem ainda seis apêndices, que tratam de vários assuntos técnicos, ligados à teoria desenvolvida. Entre eles, estão a fórmula de Taylor, a transformação de Fourier, a inversão de Möbius e a transformação de Legendre.

O tratamento geral de assuntos de Estatística Teórica certamente não é fácil, e o livro de Barndorff-Nielsen exige do leitor uma boa porção de paciência, e vontade de consultar outras fontes de informação, além de um bom fundamento em vários campos da Matemática. Mas, para quem se interessa pelos fundamentos da Estatística, o livro é indispensável.

Bent Jørgensen, IMPA

INSTRUÇÕES PARA SUBMISSÃO DE ARTIGOS À RBEs

Os artigos remetidos para publicação deverão ser submetidos em 3 vias (que não serão devolvidas) para:

Djalma G.C.Pessoa
Editor Responsável - RBEs
ENCE
Rua André Cavalcanti, l06
Bairro de Fátima
20231 - Rio de Janeiro - RJ

– Os artigos submetidos à RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.

– Cada autor receberá, gratuitamente, 20 separatas de seu artigo.

Instruções para preparo de originais:

1. O texto deve ser datilografado em papel branco tamanho ofício, em um só lado, em espaço duplo, com margem de 3 cm em todos os lados do papel, sem rasuras ou emendas que dificultem sua leitura e compreensão. As páginas devem ser numeradas seqüencialmente, contendo até 30 linhas de 72 batidas cada. Os autores interessados poderão solicitar ao Editor responsável laudas-padrão. Todas as cópias submetidas devem ser legíveis.

2. A primeira página do original (folha de rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se para cada um a filiaç˜ao e endereço permanente para correspondência. Agradecimentos a colaboradores, outras instituições e auxílios recebidos devem figurar também nesta página.

3. A segunda página do original deve conter resumos em português e em inglês (Abstract) destacando os pontos relevantes do artigo. Cada resumo deve ser datilografado seguindo o mesmo padrão do restante do texto, em um único parágrafo, sem fórmulas, com no máximo 150 palavras.

4. O artigo deve ser dividido em seções numeradas progressivamente, com títulos concisos e apropriados. Subseções, todas, inclusive a primeira, devem ser numeradas e receber título apropriado.

5. Sentenças ou palavras entre parênteses, sublinhadas ou em tipos diferentes (itálico) não devem ser usadas. Notas de rodapé devem ser evitadas. Abreviações e siglas tais como i.i.d, g.l., A.A.S, ANOVA e símbolos especiais tais como $\forall$ e $\xrightarrow{\mathcal{D}}$ não devem ser empregados.

6. A citação de referências no texto deve ser feita de acordo com os exemplos apresentados a seguir:

a) Pereira e Portugal (1987);

b) Costa (1987), p. 39); e

c) Hansen et alii (1953, cap. 5).

As referências listadas ao final devem corresponder exatamente às que foram citadas no texto. Referências a documentos não publicados pressupõem que o autor poderá fornecer cópia do material citado. A listagem final das referências deve ser apresentada em ordem alfabética do último sobrenome do autor, e para um mesmo autor, em ordem cronológica de publicação, observada a Norma NB-66/78 (ABNT).

7. As tabelas e gráficos devem ser apresentados em folhas separadas, precedidas de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas seqüencialmente (Tabela 1, Figura 3, etc.) e referidas nos locais de inserção pelos respectivos números. Quando houver tabelas e demonstrações extensas ou outros elementos de suporte, podem ser empregados apêndices. Os apêndices devem ter título e numeração tal como as demais seções do trabalho.

8. As fórmulas matemáticas devem ser apresentadas com clareza para evitar problemas de interpretação. Siga as regras indicadas e os exemplos mostrados.

a) arranje os parênteses em ordem { [( )] };

b) siga as convenções usuais para e, exp, log, etc;

c) prefira a forma $x^{1/2}$; nunca use $\sqrt{x}$;

d) notações tais como $\hat{\tilde{x}}$, ou $\underline{x}$ ou $(x + y)$ nunca devem ser usadas;

e) sub e superíndices (inclusive de segunda ordem) devem ser, se possível, alinhados horizontalmente, ou do contrário claramente marcados em tinta de cor; evite sub e superíndices de ordem maior que 2;

f) use sigma grego maiúsculo Σ apenas para indicar somatórios;

g) procure fazer distinção clara entre caracteres que se confundem facilmente, tais como:

w (dáblio) e ω (ômega minúsculo)
v (vê) e ν (ni minúsculo)
o (o minúsculo), O (o maiúsculo) e 0 (zero)
1 (um) e l (ele minúsculo)
h) procure usar a notação indicada nos exemplos a seguir:

Use	**Não Use**
$\mathrm{var}(x)$	var$x$ ou VAR(x)
cov	Cov ou COV
pr para probabilidade	P ou Pr
tr para traço de matriz	Traço ou traço
$E(X)$ para esperança de X	EX ou $\mathcal{E}(X)$
$\log(x)$	$\log_e(x)$ ou $ln(x)$
$a/(bc)$ ou $a(bc)^{-1}$	a/bc
0,2	0.2 ou .2 ou ,2
1.000.00	1000 ou 1000,00
$x_1, \ldots, x_n$	$x_1, x_2, \ldots, x_n$
$\Gamma\left(\frac{1}{2}n + \frac{1}{2}\right)$	$\Gamma[(n+1)/2]$ ou $\Gamma\binom{n+1}{2}$
ab	$a \cdot b$ ou $a \times b$

9. Equações devem ser numeradas somente se citadas no texto; nesse caso, a numeração deve ser dada entre colchetes [ ] e ser colocada alinhada junto à margem direita. Expressões ou equações longas e/ou importantes devem ser destacadas, isto é, apresentadas em linhas separadas. Fórmulas curtas devem ser deixadas no texto para poupar espaço, quando possível. Nenhuma fórmula deixada no texto deve ter mais de uma linha de altura. Portanto $\Sigma_{i=1}^{n} x_i$ não deve ser deixada no texto, ou então deve ser escrita Σx_i (se os limites de somação forem óbvios). A forma Σ_1^n não deve ser usada sob qualquer hipótese. Da mesma forma $\binom{a}{b}$ não deve ser deixada no texto. Equações ou expressões muito longas devem ser evitadas, sempre que possível, introduzindo-se notação apropriada.

10. Gráficos e diagramas para publicação devem ser traçados em papel branco, com nitidez e boa qualidade, para permitir que a redução seja feita mantendo qualidade. Fotocópias não serão aceitas. É fundamental que não existam erros quer no desenho quer nas legendas ou títulos.

ENTRE EM CONTATO COM O IBGE
FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA

SEDE — Presidência

Av. Franklin Roosevelt, 166
20021 — RIO DE JANEIRO — RJ — Telefone: 220-9442

CDDI — Centro de Documentação e Disseminação de Informações

Av. Beira Mar, 436
20021 — RIO DE JANEIRO — RJ — Telefone: 220-0511

BRASÍLIA

SCS — Quadra 06 — Bloco "A"
70300 — BRASÍLIA — DF — Telefone: 224-6998

UNIDADES REGIONAIS NAS CAPITAIS
ENDEREÇO E TELEFONE

AC/RIO BRANCO — Rua Benjamin Constant, 506, Tel.: 224-1540
AL/MACEIÓ — Rua Tibúrcio Valeriano, 125, Tel.: 223-5088
AM/MANAUS — Rua Lobo D'Almada, 272, Tel.: 232-0152
AP/MACAPÁ — Av. Antonio Coelho de Carvalho, 301, Tel.: 222-2796
BA/SALVADOR — Av. Estados Unidos, 50, Tel.: 243-9277
CE/FORTALEZA — Rua Major Facundo, 733, Tel.: 231-5352
ES/VITÓRIA — Rua Duque de Caxias, 267, Tel.: 222-5004
GO/GOIÂNIA — Av. Tocantins, 675, Tel.: 223-3307
MA/SÃO LUÍS — Rua Joaquim Távora, 49, Tel.: 222-0350
MT/CUIABÁ — Av. XV de Novembro, 235, Tel.: 322-2121
MS/CAMPO GRANDE — Rua Barão do Rio Branco, 1431, Tel.: 721-1902
MG/BELO HORIZONTE — Rua Oliveira, 523, Tel.: 223-0554
PA/BELÉM — Av. Gentil Bittencourt, 418, Tel.: 222-7195
PE/RECIFE — Rua do Hospício, 387, Tel.: 231-0811
PB/JOÃO PESSOA — Rua Irineu Pinto, 94, Tel.: 241-1560
PI/TERESINA — Rua Simplício Mendes, 436, Tel.: 222-4161
PR/CURITIBA — Rua Carlos de Carvalho, 552, Tel.: 234-9122
RJ/RIO DE JANEIRO — Rua Humaitá, 85, Tel.: 286-2672
RN/NATAL — Praça Pedro Velho, 435, Tel.: 222-3695
RO/PORTO VELHO — Av. Duque de Caxias, 1223, Tel.: 221-5143
RR/BOA VISTA — Av. Getúlio Vargas, 76-E, Tel.: 224-4425
RS/PORTO ALEGRE — Av. Augusto de Carvalho, 1205, Tel.: 228-6444
SC/FLORIANÓPOLIS — Rua João Pinto, 12, Tel.: 222-0733
SE/ARACAJU — Rua Riachuelo, 1017, Tel.: 222-8197
SP/SÃO PAULO — Rua Urussuí, 93, Tel.: 883-0077