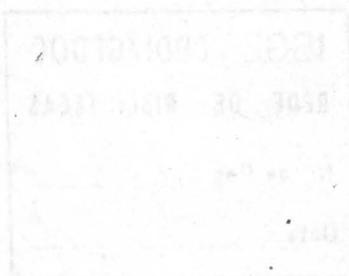


INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATISTICA — IBGE

ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS — ENCE



RELATÓRIOS TÉCNICOS

Nº 02/88

USO DE AMOSTRAGEM EM SIMULAÇÃO
DE
LEGISLAÇÃO TRIBUTÁRIA

José Carlos da Rocha C. Pinheiro (ENCE/IBGE)
Manuel Martins Filho (DISUL/SERPRO)

31(047.3)

R382r

2/88

GE 9014.6

IBGE - CDDI/GEDOC
REDE DE BIBLIOTECAS
N.º de Reg: <u>198</u>
Data: <u>25/01/90</u>

SUMARIO

- 1 - O porque da Amostra do IRPF
- 2 - Histórico
 - 1974 - Dimensionamento com Estatísticas Básicas
 - 1976 - Uso da amostra velha para dimensionar amostra nova
 - 1980 - Introdução do conceito de Especiais
- 3 - Metodologia
 - 3.1 - Estratificação
 - 3.1.a Variáveis de Estratificação
 - 3.1.b Definição dos Especiais
 - 3.1.c Construção dos Estratos
 - 3.2 - Alocação
 - 3.2.a Metodologia
 - 3.2.b Resultados da Alocação
 - 3.2.c Validação Preliminar
 - 3.3 - Seleção da Amostra
 - 3.3.a Programa Amostrador
 - 3.3.b Procedimento Amostral
 - 3.3.c Estatísticas do Universo
- 4 - Validação da Amostra
 - 4.1 - Erros Relativos
 - 4.2 - Ganhos associados à Estratificação
- 5 - Conclusões
- 6 - Bibliografia

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 SERPRO desenvolveu o Sistema Simulador de Legislações, que vem sendo usado ao longo dos últimos doze 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.

1 - O PORQUE 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 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 de forma rápida e flexível, medir 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 a 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 sua declaração é refeita a 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 apresentado, 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(Especiais).

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 à 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 frequência por classe de renda da quase totalidade das variáveis envolvidas com o tributo.

Através de aproximações, calculou-se por classe de renda , as médias e as variâncias das variáveis envolvidas no dimensionamento

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, oferecia ao usuário um módulo para alocação de Amostras conversacional, e 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 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 adquirida no período de 1975 até 1979 com o dimensionamento, coleta e uso de Amostras, o SERPRO desenvolveu o 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 a 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 sub-populações 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.
- o interesse em obterem-se estimativas de precisão pré-determinada para subdivisões da População.
- 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 sub-populações homogêneas. A obtenção de estimativas de 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.a 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 frequência 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 o 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)⁽¹⁾. Foram obtidas, ao todo, nove faixas de Renda Tributável para Estratificação dos declarantes.⁽²⁾

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 (A1)
- Imposto sobre Lucro na Alienação de Participação Societária (A2)
- Imposto sobre Lucro na Alienação de Imóveis (A3)
- Variação de Bens (A4)
- Rend. Trib. Exclusivamente na Fonte (A5)

(1) Na determinação das especificações necessárias à seleção da Amostra de declarantes referente ao exercício de um certo ano, faz-se sempre uso da Amostra coletada para o exercício anterior.

(2) Na seção 3.1.c é apresentada a metodologia utilizada para construção dos Estratos de Renda Tributável. Os limites das diferentes faixas são ali apresentados.

A cada uma das variáveis supracitadas associou-se um ponto de corte (C_1 a C_5), que deixa acima de si os valores mais expressivos da População⁽¹⁾. 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:

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

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

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 9 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.e. Não-Especiais através das 9 classes de Renda Tributável) procedeu-se a uma Amostragem Binomial sem reposição.⁽²⁾

(1) Na seção 3.1.b é apresentada a metodologia utilizada para determinação dos pontos de corte.

(2) No seção 3.3 é apresentado o algoritmo de seleção da Amostra nos Estratos de Renda Tributável dos Não-Especiais.

3.1.b 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 cinco variáveis de classificação (Renda Não-Tributável(A1), Participação Societária(A2), Lucro Imobiliário(A3), Variação de Bens(A4) e Rendimentos Tributados Exclusivamente na Fonte(A5)).

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 ;
- o estabelecimento de pesos relativos para as variáveis de classificação na formação do grupo dos Especiais (i.e. 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:

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

$$(ii) \quad PC (A_i \geq C_i) = P_i \cdot \delta, \quad i=1,2,\dots,5$$

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).⁽¹⁾

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

- (i) a especificação das distribuições das variáveis de classificação.
- (ii) a determinação do δ que satisfaça o sistema de equações anterior.

(1) De fato C_i é o percentil de ordem $1 - P_i \cdot \delta$ da distribuição de A_i

(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 frequências (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 seguem uma distribuição LOG-NORMAL.

Em outras palavras, o logaritmo das restrições positivas das variáveis de classificação, se distribuem segundo uma Normal. Representamos a hipótese assumida por :

$$\text{LOG} (A_i^*) \sim N (\mu_i , \sigma_i^2) , i = 1, 2, \dots, 5$$

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 $\text{LOG}(A_i^*)$ respectivamente).

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

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

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 desconhecidas e devem ser substituídos na fórmula acima por estimativas.

Observe-se, assim, que os pontos de corte obtidos são, na verdade, 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.

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

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

Para ver isto note inicialmente que :

$$\forall i=1,2,\dots,5 \quad P(A_i \geq C_i) = P_i \cdot \delta \leq 1$$

$$\Rightarrow \forall i=1,2,\dots,5 \quad \delta \leq 1 / P_i$$

$$\Rightarrow \delta \leq \min (\alpha/P_1, \alpha/P_2, \alpha/P_3, \alpha/P_4, \alpha/P_5) = \alpha/P_m$$

Por outro lado,

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

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

$$b) \text{ a função } g(\delta) = P \left(\bigcup_{i=1}^5 (A_i \geq C_i(\delta)) \right)$$

é contínua e crescente no intervalo $\left[\alpha / \sum_{i=1}^5 P_i, \alpha/P_m \right]$

$$\text{com } g \left(\alpha / \sum_{i=1}^5 P_i \right) \leq \alpha \text{ e } g(\alpha/P_m) \geq \alpha$$

Para ver isto, note que :

$$g(\delta) = 1 - P \left(\bigcap_{i=1}^5 (A_i < C_i(\delta)) \right) =$$

$$1 - F (C_1(\delta), C_2(\delta), C_3(\delta), C_4(\delta), C_5(\delta))$$

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 Normalidade da distribuição das restrições positivas das variáveis de classificação.

As duas observações apresentadas acima implicam na 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

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

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).

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

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

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

$$\delta = \delta_n^*$$

onde ϵ 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 apresentado a seguir:

$$i) a_0 = \alpha / \sum_{i=1}^5 P_i ; b_0 = \alpha / P_m ; \delta_0 = (a_0 + b_0) / 2$$

ii) se $g(\delta_n) > \alpha$ então :

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

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

se $g(\delta_n) < \alpha$, então :

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

$$a_{n+1} = \delta_n ; b_{n+1} = b_n .$$

Os pontos de corte adotados para definição dos Especiais serão:

$$\hat{C}_i = \exp (\hat{\mu}_i + \hat{\sigma}_i \cdot Z(1 - P_i \cdot \delta^* / \gamma_i)) , i=1,2,\dots,5$$

Observe-se que os valores obtidos, como se baseiam em uma Amostra de declarantes do exercício imediatamente anterior aquele 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 :

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

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 :

$$\alpha = 16.000/7.000.000 = 2.286 \times 10^{-9}$$

$$P_1 = P_5 = 2 ;$$

$$P_2 = P_4 = 1.1 ;$$

$$P_3 = 1 ;$$

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 de δ necessariamente está contido no intervalo :

$$[0.0023/(2+2+1.1+1.1+1) , 0.002286/2] = [3.194 , 11.43] \times 10^{-4}$$

De fato, obteve-se ao final dos procedimentos numéricos.

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

Os valores encontrados para os pontos de corte, já com valores corrigidos para a inflação do período⁽¹⁾, são apresentados na tabela abaixo.

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

Na tabela acima n_i representa o número de indivíduos na População com valores de A_i acima do ponto de corte C_i . Note-se que a soma dos n_i não é igual ao número de Especiais.

(1) Utilizou-se para correção o INPC acumulado referente ao ano de 1985

3.1.c 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-se outra variável, correlacionada com a ideal, que esteja a disposição para os cálculos necessários. Face à 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), pgs 129 e 130, seguindo trabalho de DALENIUS e HODGES (1959). Esta metodologia, que denominaremos Método das Frequências, foi implementada no SANDA, fazendo parte de sua rotina de Estratificação.

O Método das Frequências 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 = no. de indivíduos do estrato h

$W_h = N_h/N$, 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

$\bar{y}_{st} = \sum_{h=1}^L W_h \bar{y}_h$, o estimador da média da característica y na População , baseado na Amostra Estratificada

Então , sob alocação de NEYMAN (i. e. $n_h = n \cdot (W_h \cdot S_h) / (\sum_h W_h \cdot S_h)$)

temos :

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

Se a População for suficientemente grande para ignorar-se o fator de correção para Populações finitas, temos :

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

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

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

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 que :

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

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

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

Decorre então que :

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

Fazendo y

$$Z(y) = \int_{y_0}^{y_h} \sqrt{f(t)} dt, \quad \text{então}$$

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

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 contraindica 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 tão mais crítico quanto maior a quantidade de Estratos.

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

- foram obtidos os limites ótimos, segundo o Método das Frequências exposto no item anterior, para uma grande variedade de números 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 .

- 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 abaixo apresenta a variação dos erros relativos da variável de Estratificação para as diversas quantidades de Estratos consideradas.

NÚMERO DE ESTRATOS

ERRO RELATIVO (x10e-1)	01	02	03	04	05	06	07	08	09	10
	10.51	5.76	4.13	3.33	2.65	2.21	1.98	1.72	1.54	1.42

ERRO RELATIVO (x10e-1)	11	12	13	14	15	16	17	18	19	20
	1.34	1.22	1.15	1.09	1.03	1.01	0.95	0.92	0.88	0.80

Analisando-se a tabela acima, verifica-se que a partir de 9 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 Frequências, já corrigidos para cruzeiros de 1985, são apresentados abaixo :

ESTRATO	LIMITES (CR\$ 1985)	
01	0	- 12.983.750
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	

3.2 ALOCAÇÃO

3.2.a 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 que compõem o Estrato dos Especiais são selecionados com Probabilidade 1. O restante da Amostra é distribuído, entre os Não-Especiais através 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 denotando-se por :

N_h - o número de indivíduos do estrato h

Y_{hi} - o valor da característica sob investigação para o i-ésimo indivíduo do estrato h

\bar{Y}_h - média da característica no estrato h

S_h^2 - variância no estrato h $(= \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 / (N_h - 1))$

n_h - tamanho da Amostra no estrato h

f_h - fração amostral no estrato h $(= n_h / N_h)$

N - número de indivíduos na População

W_h - peso do estrato h $(= N_h / N)$

\bar{Y} - média populacional da característica

\bar{y}_h - estimador de \bar{Y}_h

\bar{y}_{st} - estimador de \bar{Y} baseado na estratificação $(= \sum_{h=1}^L W_h \bar{y}_h)$

então, para um determinado tamanho total de Amostra n, tem-se a alocação de NEYMAN dada por :

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

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 um conjunto delas. Desta forma é desaconselhável a escolha de uma única variável para fins de alocação, já que isso poderia implicar na 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⁽¹⁾, consideradas mais importantes. A alocação final é obtida a partir da combinação destas alocações segundo o esquema abaixo.

Sejam :

$nh(p)$ - o tamanho da Amostra no estrato h , com base na alocação de NEYMAN para a variável p

nh - o maior tamanho de Amostra para o estrato h
($= \max (nh(p))$)
 p

então, a alocação final é dada por :

$$nh = \left(nh' / \left(\sum_{h=1}^L nh' \right) \right) \cdot n$$

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

(1) Foram utilizadas além das variáveis já mencionadas na seção de Estratificação, a saber: Rendimento Tributável, Rendimento Não-Tributável, Rendimento Tributado Exclusivamente na Fonte, Imposto sobre Lucro na Alienação de Participações Societárias, Imposto sobre Lucro na Alienação Imóveis, Variação de Bens, as seguintes variáveis de interesse: Rendimento Categoria-G, Abatimento Médico, Renda Líquida, Imposto Líquido Devido e Redução Investimento.

3.2.b 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 referente ao Estrato de 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 progressa , que um tamanho esperado de 15.000 Especiais seria suficiente.

O restante da Amostra (45.000 declarantes) foi alocado aos Estratos segundo a metodologia descrita no item anterior, tendo-se obtido :

ESTRATO	Nh	nh	$fh(x10^{-2})$	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

3.2.c VALIDAÇÃO PRELIMINAR

A partir da Amostra do exercício anterior, foi feita uma validação preliminar da Amostra obtida. As variâncias dos estimadores das Médias 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
- c- com a estratificação completa

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 na situação a
- Var(b) - Variância na situação b
- Var(c) - Variância na situação c
- $G_1 = \text{Var}(a)/\text{Var}(b)$
- $G_2 = \text{Var}(b)/\text{Var}(c)$

VARIÁVEL	VAR(a) 1e+22	VAR(b) 1e+22	VAR(c) 1e+22	G1	G2
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	6,08xE-5	5,96xE-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

Observe-se que as variáveis positivamente correlacionadas com a Renda Tributável apresentam ganhos consideráveis devido a 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 variáveis analisadas apresentaram coeficientes de variação bastante reduzidos , tendo-se obtido as seguintes estimativas :

VARIÁVEL	Coef. de Variação ($\times 10^{-2}$)
RENDA TRIBUTÁVEL	0,0699
RENDA LÍQUIDA	0,1690
IMPOSTO LIQ. DEVIDO	0,2930
RENDA NÃO-TRIB	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

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.a 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.

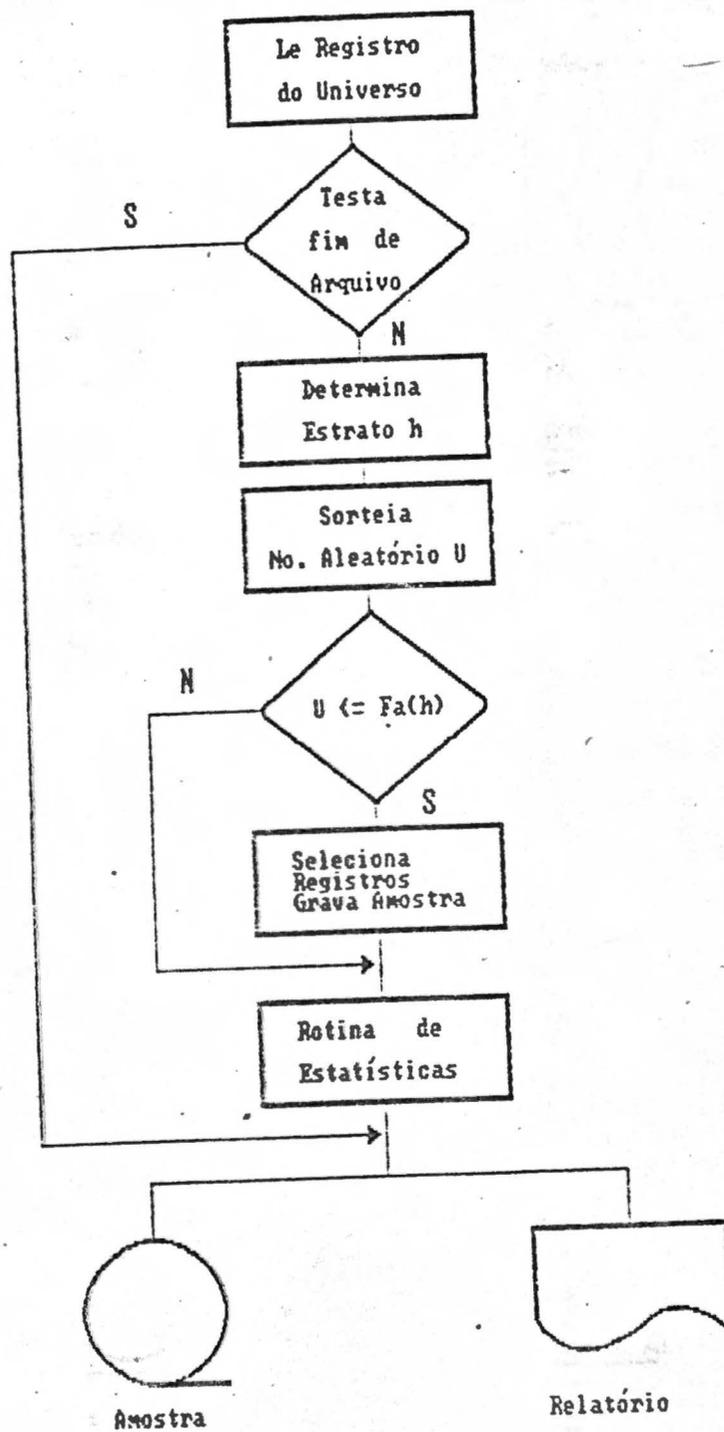
Ao programa Amostrador são fornecidas as seguintes informações :

- Limites dos Estratos de Renda Tributável
- Definição dos Estratos dos Especiais
- Frações Amostrais (FA) (por Estrato)
- 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 $U \leq FA(h)$ seleciona o declarante para a Amostra (grava registro no arquivo Amostra)
- 5- Processa Rotina de Estatísticas do Universo.
- 6- Repete o procedimento até o final do arquivo de entrada (Universo de Declarantes do IRPF)

Fluxograma Simplificado



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

3.3.b 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 sequencialmente 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 α daquele Estrato.

Ou seja, se para $U \leq FA(h)$ o indivíduo do Estrato h é selecionado, então o Estrato h terá uma fração amostral aproximadamente igual a α . 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 $FA'(h)$, $h=1,2,\dots,L$ próximas das frações amostrais teóricas $FA(h)$, $h=1,2,\dots,L$ desejadas. (1)

(1) As frações amostrais Fh' 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 Fh' encontre-se bem próxima da fração teórica Fh .

As diferenças entre as frações amostrais teóricas e reais são exibidas na tabela abaixo.

ESTRATO	FRAÇÃO AMOSTRAL		ERRO RELATIVO $\frac{(Fa-Fa')}{Fa}$
	TEÓRICA(Fa)	REAL(Fa')	
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

max=2.09% min=0.22%

TABELA-1

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 ($8,33 \times E+6$ declarantes).

Dado o exposto, optou-se pelo procedimento de seleção que procura manter as frações amostrais teóricas especificadas na alocação. Observando a Tabela-1 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 (~ 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 contribuintes 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_h , $h=1,2,\dots,L$ 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 3 anos consecutivos(83,84, e 85).

Tal procedimento permitiu estudos mais detalhados (precisos) sobre os impactos produzidos por alterações na Legislação do IRPF , principalmente no que se refere a 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-2) contém o resultado do dimensionamento obtido à partir da Amostra-85 (Ver seção- 3.2).

ESTRATO	Nh	nh	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

Tabela-2

O tamanho total da Amostra dimensionada (60.982) resulta em uma fração amostral global de 0,00857 (0,857%).

Na Tabela-3 estão as medidas obtidas na seleção da Amostra-86.

ESTRATO	Nh	nh	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

Tabela-3

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-1) 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.c 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
- Variação Patrimonial

Os totais serão usados na fase de validação da Amostra , onde serão comparados os valores obtidos no Universo com os valores estimados à 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 a Estratificação e a 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 :

$$e(\hat{\theta}) = \left(\frac{\hat{\theta} - \theta}{\theta} \right) \cdot 100$$

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 abaixo.

ERROS RELATIVOS DO TOTAL POR ESTRATO

RENDG	RTRIB	AMED	RLIQ	REDINV	I L D	PSOCIE	LIMOV	RNTRIB	RTXF	V
11.57	0.05	3.03	2.81	-28.76	20.49	*	36.79	-2.99	14.50	-
-11.92	0.06	-4.11	0.01	8.55	-1.47	418.47	90.12	-7.03	-17.25	-
-3.77	-0.07	2.80	-0.95	-5.10	-2.83	52.54	-82.63	-0.74	-6.00	-
-18.64	-0.07	-1.01	-0.21	2.63	-0.50	*	118.01	-5.76	13.95	-
-4.49	-0.12	-2.93	-0.02	-0.23	0.02	*	154.23	-3.41	5.64	-
-3.11	-0.28	-1.63	-0.02	1.80	-0.12	8.37	-15.76	0.93	2.78	-
6.85	0.08	-0.96	0.23	0.35	0.35	-27.97	-2.25	-1.22	-8.19	-
6.87	-0.18	1.44	-0.18	1.57	-0.24	-51.01	24.67	0.91	-2.66	-
-0.32	-0.02	2.45	-0.01	0.26	-0.11	-16.79	10.36	1.20	-0.60	-
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	-

ERROS RELATIVOS DO C. V. POR ESTRATO

RENDG	RTRIB	AMED	RLIQ	REDINV	I L D	PSOCIE	LIMOV	RNTRIB	RTXF	V
-5.62	0.72	-39.17	-58.39	-15.35	6.20	*	-26.28	-14.23	2.46	-1
5.69	2.05	-48.85	-13.28	-8.46	-1.777	-61.13	-21.65	-7.23	2.77	-11
2.27	3.60	-22.26	-13.63	1.78	1.12	-61.45	-40.60	0.77	35.50	13
9.34	0.84	-30.26	-2.39	-2.49	-1.511	*	-30.66	-5.80	1.53	9
0.23	6.10	-10.02	-4.24	0.18	0.624	*	-37.88	-4.41	1.27	-10
0.79	5.98	-18.05	-3.52	-1.09	0.40	5.36	-4.94	5.42	-2.64	2
-0.42	4.64	-5.51	-0.42	-1.60	-0.18	-4.577	5.76	-2.86	6.70	-3
-2.18	4.61	9.88	-6.78	-1.86	1.43	-8.02	-9.56	-0.74	2.68	2
-0.81	-1.82	70.38	-1.01	-1.24	-0.83	11.18	-2.01	0.07	-1.32	-0
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
-0.56	1.70	-19.91	-14.97	-1.20	-0.11	-1.82	12.30	-4.81	9.05	-0

(1) 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.

(2) Os erros relativos para os Especiais são nulos, já que todos indivíduos deste grupo são selecionados na Amostra.

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 mais alta) e visou-se a 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 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 de 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:

- V₁ - Variância com A. A. S. S. R;
- V₂ - Variância com estratificação pelos Especiais;
- V₃ - Variância com a estratificação por Renda Tributável;
- V₄ - Variância com estratificação cruzada.

No cálculo das Variâncias acima, utilizou-se sempre o mesmo tamanho de Amostra, tendo sido necessário, no caso de V₃, refazer a alocação para distribuir, entre os Estratos de Renda Tributável, a parcela da Amostra representada pelos Especiais. As duas alocações são apresentadas abaixo:

ESTRATO	ANTIGA	NOVA
01	4476	7093
02	6927	6795
03	2944	5089
04	3215	5234
05	2792	4251
06	6738	10298
07	4816	9091
08	4259	5570
09	20210	25657
ESPECIAIS	22599	0
TOTAL	78976	78976

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 Rend. Trib.)
 $G_5 = V_3/V_4$ (Ganho adicional da estratificação por Cond. Especial)

A tabela abaixo apresenta os ganhos obtidos para as variáveis alvo.

Var. / Ganhos	G1	G2	G3	G4	G5
REND. CEDULA-G	2.05	2.83	2.18	1.06	0.77
REND. TRIB.	0.58	42.14	48.21	83.27	1.14
ABAT. MEDICO	0.84	1.30	1.12	1.33	0.86
RENDA LIQUIDA	0.74	13.66	11.68	15.79	0.85
REDUÇÃO INVEST.	0.72	4.45	2.86	3.99	0.64
IMP. LIQ. DEVIDO	1.38	33.94	51.25	37.01	1.51
IMP. LUCRO PSOCIE	40.5E+4	8.56	35.8E+4	0.88	41.8E+3
IMP. LUCRO IMOB.	31.50	1.26	28.70	0.91	22.84
REND. NAO 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

O ganho G_3 é o mais relevante para julgamento da qualidade da Amostra. Por ele pode-se concluir pela exelência 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 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 influenciando 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 ($\sim 8.9 \times 10^6$) 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.

6 - BIBLIOGRAFIA

- Cochran, W. , "Sampling Techniques" , Third Edition - (1977)
John Wiley & Sons Inc.
- Dalenius, T. , Hodges, J.L.Jr. , (1959)
Minimum Variance Stratification.
Journal of the American Stat. Association (JASA) 54, pg 88-101
- Xavier, A. , Pinheiro, J.I. , Azevedo, L.M. , Martins, M. ,
Simulador de Legislações para o Imposto de Renda
Pessoa Física-(IRPF) em atas do II SINAPE - Campinas-1976