INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS - ENCE

R E L A T Ó R I O S T É C N I C O S

Nº 004/88

AVALIAÇÃO DOS EFEITOS DE REDUÇÃO DA FRAÇÃO DE AMOSTRAGEM NO CENSO DEMOGRÁFICO

José Carlos da Rocha C. Pinheiro - ENCE
José Matias de Lima - DPE/NME

RIO DE JANEIRO
Dezembro/1988

SUMÁRIO

1 - Introdução

2 - Metodologia

3 - Análise e Apresentação dos Resultados

4 - Conclusões

5 - Anexo

6 - Bibliografia

RESUMO

Neste trabalho é desenvolvida uma medida de perda de eficiência da precisão das estimativas ,decorrente da redução da fração de amostragem .

A medida ,aqui apresentada, foi desenvolvida no sentido de suprir as deficiências da medida clássica de eficiência relativa entre estimadores , para atender aos propósitos do trabalho, ou seja, analisar até que ponto a perda de eficiência representada pela redução da fração de amostragem poderá comprometer a qualidade das estimativas obtidas através da amostra.

É também desenvolvido um estudo referente à fração de amostragem necessária para garantir um nível pré- fixado de precisão na estimação de características de interesse.

Uma aplicação da metodologia desenvolvida é incluída,tomando-se por base as características de domicílios investigas no Censo Demográfico de 1980 que tiveram suas estimativas divulgadas à nível de município e as frações de amostragem que vêm sendo cogitadas para utilização no Censo Demográfico de 1990.

1. INTRODUÇÃO

Este documento pretende fornecer subsídios para o enriquecimento das discussões referentes aos estudos de redução da amostra do Censo Demográfico de 1990.

O principal objetivo deste estudo consiste em avaliar e analisar a perda de eficiência das estimativas de total de características de domicílios obtidas com a utilização de uma fração de amostragem inferior à adotada no Censo de 1980 [1], e, também, avaliar as frações amostrais necessárias para atender um determinado nível de precisão fixado à priori.

Costa, L. N. & Lima, J. M. [2], constataram, através de amostras com fração de 1/10, simuladas a partir da amostra de 25% do Censo Demográfico de 1980 e considerando três municípios cuja população não excedia 20 000 habitantes, que a amostra de 10% fornece estimativas de características de domicílios, a nível de município, sem grande perda de precisão comparado com aquelas obtidas na amostra de 25% do Censo Demográfico de 1980.

Neste trabalho, foram considerados 48 municípios abrangendo todas as regiões brasileiras com contingentes populacionais variando de 1966 a 1 307 608 habitantes. A exemplo de Costa, L.N. & Lima, J. M [2], foram utilizadas no estudo as variáveis referentes às características de domicílios do plano tabular do Censo Demográfico de 1980 que tiveram suas estimativas divulgadas à nível de município. A relação dos municípios bem como das variáveis consideradas encontram-se nos anexos 1 e 2, respectivamente.

É desenvolvida neste trabalho uma medida de perda de eficiência da precisão das estimativas associada à redução da fração de amostragem. Esta medida é definida a partir das perdas de eficiência individuais do elenco de variáveis considerado, calculadas através da precisão das estimativas.

São esperadas maiores perdas de eficiências e também frações amostrais elevadas para

atender determinados níveis de precisão previamente estipulado, quando considerados municípios com baixo contingente populacional.

Na segunda seção apresenta-se a metodologia utilizada no trabalho. Os métodos ali apresentados são descritos, propositalmente, de forma teórica sem restrições a este estudo particular, de modo a permitir sua utilização e/ou adaptação a outros estudos de natureza semelhante.

Na terceira seção são apresentados e analisados os resultados da aplicação da metodologia ao conjunto de municípios considerados. A análise se desenvolve através de gráficos com o intuito de facilitar o seu entendimento.

Finalmente são apresentadas algumas conclusões bem como sugestões sobre os resultados obtidos.

2.METODOLOGIA

O objetivo principal deste trabalho, como já mencionado anteriormente, é o de avaliar os efeitos da redução da fração amostral utilizada no Censo Demográfico do IBGE sobre a precisão dos estimadores das características investigadas a partir da amostra. Ou seja, pretende-se analisar até que ponto a perda de eficiência representada pela redução da fração amostral poderá comprometer a qualidade das estimativas a serem divulgadas.

a) Medidas de Eficiência

A medida clássica de eficiência relativa entre estimadores é definida a partir do quociente entre os erros quadrados médios dos mesmos. Assim, se θ é um parâmetro de interesse e $\hat{\theta_1}$ e $\hat{\theta_2}$ dois estimadores alternativos, a eficiência de $\hat{\theta_1}$ com relação à $\hat{\theta_2}$ é definida como segue (ver [3], por exemplo).

$$e_\theta(\hat{\theta_1}, \hat{\theta_2}) = \frac{EQM_\theta(\hat{\theta_2})}{EQM_\theta(\hat{\theta_1})} \tag{1}$$

Neste trabalho, os estimadores considerados diferem entre si apenas pela fração amostral utilizada no seu cálculo. Pode-se, então, usar a seguinte notação, para caracterizar a dependência com relação à fração amostral (f):

$$\hat{\theta} = \hat{\theta}(f) \tag{2}$$

De maneira análoga à anterior, definimos a eficiência entre duas frações amostrais f_1 e f_2, na estimação de um parâmetro de interesse θ, da seguinte forma:

$$e_\theta(f_1, f_2) = \frac{EQM_\theta(\hat{\theta}(f_2))}{EQM_\theta(\hat{\theta}(f_1))} \tag{3}$$

Como o parâmetro de interesse, θ, é o mesmo nas duas frações consideradas, pode-se, na ausência de vício dos estimadores e supondo $\theta \neq 0$, reescrever a fórmula acima em termos dos coeficientes de variação (CV) dos estimadores envolvidos:

$$e_\theta(f_1, f_2) = \left(\frac{CV(\hat{\theta}(f_2))}{CV(\hat{\theta}(f_1))} \right)^2 \tag{4}$$

Ocorre que esta medida clássica não é interessante para comparar frações amostrais, já que:

(i) ela mede a eficiência com relação a uma única característica (quando o que se deseja é avaliar o impacto de redução para o conjunto de variáveis investigadas); e

(ii) ela não leva em conta a precisão absoluta dos estimadores (um aumento do CV de 0,5% para 1% é, por esta medida, considerado tão importante quanto o acréscimo de 5% para 10%, ou 10% para 20%, enquanto que os impactos sobre a qualidade das estimativas são certamente distintos, nos três casos mencionados).

A fim de ilustrar os pontos mencionados no parágrafo anterior, considere-se, por exemplo, um esquema de amostragem aleatória simples sem reposição (AASSR). Na estimação de médias ou totais populacionais, tem-se (ver [4], por exemplo)

$$e_\theta(f_1, f_2) = \frac{N^2 S^2 (1 - f_2)/n_2}{N^2 S^2 (1 - f_1)/n_1} = \frac{f_1(1 - f_2)}{f_2(1 - f_1)} \tag{5}$$

onde:

N = tamanho da população;

S^2 = variância da característica investigada, na população;

n_i = tamanho da amostra associado à fração f_i;

Observe-se, assim, que a eficiência independe de quaisquer parâmetros da população, sendo a mesma para qualquer característica investigada e para qualquer tamanho de população.

Considere-se, por exemplo, as frações de 25% e 10%. Obtém-se, neste caso:

$$e_\theta(0,1;0,25) = \frac{0,1(0,75)}{0,9(0,25)} = 1/3$$

ou seja, há uma perda de eficiência da ordem de 67 % associada à redução da fração amostral de 25% para 10%. observe-se, ainda que, substituindo o valor encontrado acima na fórmula (4), obtém-se:

$$CV(\hat{\theta}(0,1)) = \sqrt{3} \quad CV(\hat{\theta}(0,25))$$

Note-se que esta relação é válida apenas na suposição de AASSR e na estimação de médias ou totais populacionais.

As deficiências da medida clássica de eficiência, abordadas nos parágrafos precedentes, motivaram o desenvolvimento de outra medida mais apropriada aos propósitos deste trabalho. O objetivo principal foi o de obter uma medida que:

(i) levasse em conta o conjunto de variaveis investigadas; e

(ii) dependesse, em certo sentido, do valor absoluto das precisões envolvidas.

Com relação ao (ii), o que se pretende é que a medida trate de forma diferenciada aumentos, ou decréscimos, de variabilidade, de acordo com a magnitude dos valores envolvidos. Assim, na mesma linha de raciocínio do exemplo anteriormente citado, a passagem de um CV de 0,5% para 1% não deveria ser considerada como uma perda de eficiência significativa, já que, em ambos os casos, a precisão do estimador é mais do que razoável para a maioria das aplicações práticas. De forma equivalente, a mudança do CV de 80% para

160% também não deveria ser tratada como uma perda substancial de eficiência, já que a precisão é bastante ruim nos dois casos. Restam, portanto, as situações intermediárias. Por exemplo, a passagem do CV de um estimador de 7% para 14%, com a mudança da fração amostral, poderia ser considerada uma perda significativa de eficiência, caso o limite de 10% fosse estabelecido como o máximo tolerável para a precisão dos estimadores.

Com base nas características mencionadas anteriormente, passamos, agora, a definição da medida de perda de eficiência utilizada no trabalho. Suponha que, existam **r** características sendo investigadas e que foi estabelecido um limite máximo **c**, a partir do qual o CV associado ao estimador de uma característica de interesse é considerado inaceitável, em termos da precisão requerida para os resultados a serem divulgados. Seja

$$I_c(\hat{\theta}_i) = \begin{cases} 1, & \text{se } CV(\hat{\theta}_i) > c \\ 0, & \text{se } CV(\hat{\theta}_i) \leq c \end{cases}, \qquad i = 1, \cdots, r \tag{6}$$

Define-se a medida de **perda de eficiência relativa** entre as frações f_1 e f_2 como se segue:

$$P_c(f_1, f_2) = \sum_{i=1}^{r} \mid I_c(\hat{\theta}_i(f_1)) - I_c(\hat{\theta}_i(f_2)) \mid /r \tag{7}$$

Note-se que, se, com ambas as frações, o estimador apresentar CV acima ou abaixo do limite **c**, a sua contribuição para a medida de perda de eficiência será nula. Caso haja a transposição do limite **c** quando da mudança da maior para a menor fração, o numerador da expressão anterior será acrescido de uma unidade. Observe-se, ainda, que a medida é simétrica com relação às frações f_1 e f_2, ou seja, a eficiência de f_1 com relação à f_2 é idêntica à de f_2 com relação a f_1. A fim de evitar confusão, a perda de eficiência deve ser sempre considerada como da fração menor com relação à fração maior[1].

O limite de tolerância c, cuja definição é central para o cálculo de $P_c(f_1, f_2)$, é estabelecido de forma até certo ponto subjetiva, sendo difícil chegar a um consenso sobre o valor mais adequado. Este problema pode ser contornado considerando-se a medida como uma função do limite de tolerância:

$$P(c, f_1, f_2) = P_c(f_1, f_2)$$

e obtendo-se o seu valor para diversos valores de c, dentro de uma faixa de variação que abranja os valores considerados razoáveis, em termos de aplicações práticas.

Observa-se, que, pela definição da função $P(c, f_1, f_2)$, tem-se:

$$\begin{aligned} &(i) 0 \leq p(c, f_1, f_2) \leq 1, \quad c \ \varepsilon \quad [0, +\infty) \\ &(ii) p(0, f_1, f_2) = 0, \quad f_1, f_2 \ \varepsilon \quad [0, 1) \\ &(iii) \lim_{c \to \infty} P(c, f_1, f_2) = 0 \end{aligned} \tag{8}$$

Torna-se interessante, neste ponto, introduzir o conceito de **perda de eficiência global**, que é associado a uma única fração amostral

$$P_c(f) = P(c, f, 1) = \sum_{i=1}^{r} I_c(\hat{\theta}_i(f))/r \tag{9}$$

[1]O problema poderia ser contornado eliminando-se o módulo no numerador da expressão (7) e trabalhando-se com medidas negativas de eficiência. De modo a manter intacta a intuição por trás da interpretação dos valores da medida considerada, optou-se pela definição apresentada.

Note-se que a eficiência relativa entre duas frações pode ser descrita a partir das eficiências globais, como na expressão seguinte:

$$P(c, f_1, f_2) = [P_c(f_1) - P_c(f_2)]$$

O desconhecimento dos verdadeiros coeficientes de variação faz com que, na prática, sejam utilizadas estimativas em substituição aos CVs teóricos nas fórmulas (6) e (7). Ocorre que, eventualmente, não é possível estimar o CV de um determinado estimador. Isso ocorre, por exemplo, quando nenhum indivíduo selecionado para a amostra apresenta uma determinada característica sob investigação. Duas variantes da medida anteriormente apresentada são consideradas, levando em conta este aspecto da estimação:

(i) Perda Corrigida - obtida da mesma maneira que a descrita em (7), com **r** substituído por

r' = número de características para as quais é possível obter estimativas do erro amostral;

(ii) Perda Absoluta - obtida de forma análoga à descrita em (7), fazendo-se

$$\left[I_c(\hat{\theta}(f_1)) - I_c(\hat{\theta}(f_2))\right] = 0 \quad \forall_c \quad \in \quad [0, +\infty)$$

para aquelas características para as quais não é possível estimar o erro amostral, ou seja, estas características não contribuem para a perda de eficiência. A lógica por trás desta variante de $P_c(f_1, f_2)$ é que características para as quais não se pode estimar o CV no estimador são usualmente muito rarefeitas na população(e neste caso o CV é muito elevado).

As duas medidas são utilizadas na apresentação dos resultados práticos obtidos com a aplicação da metodologia, incluídos na seção seguinte.

b) Frações Amostrais Necessárias

Com o objetivo de enriquecer a análise sobre a questão da mudança da fração amostral, incluiu-se, neste trabalho, uma parte dedicada ao estudo da fração necessária para garantir um nível pré-fixado de precisão na estimação das características de interesse.

A definição da fração necessária pressupõe a especificicação de um limite máximo de tolerância para o CV de um estimador (de forma análoga a da medida $P_c(f_1, f_2)$ e de um percentual mínimo de características que devem ser estimadas obedecendo ao limite de tolerância pré-fixado para a precisão.

Define-se a fração amostral necessária para um limite de tolerância **c** e um percentual de variáveis de interesse α da seguinte forma:

$$f(c, \alpha) = inf\{f | P_c(f) \leq 1 - \alpha\} \tag{10}$$

Para o cálculo da fração necessária, foi preciso restringir-se a um método de seleção (AASSR) e fixar uma fração de referência, fo. Observe-se que, quando o método utilizado é AASSR, a fórmula (4) é válida e obtém-se:

$$CV(\hat{\theta}(f)) = \left(\frac{fo(1-f)}{f(1-fo)}\right)^{1/2} CV(\hat{\theta}(fo)) \tag{11}$$

Sejam

$$CV(\hat{\theta}(f))_{(1)},\quad CV(\hat{\theta}(f))_{(2)},\quad CV(\hat{\theta}(f))_{(3)},\quad CV(\hat{\theta}(f))_{(4)},\quad \cdots,\quad CV(\hat{\theta}(f))_{(r)}$$

os coeficientes de variação dos estimadores das **r** características de interesse, assumindo-se a fração amostral **f**, ordenados de forma crescente. Sob AASSR, a ordenação dos CV's não é alterada pela mudança da fração amostral e obtém-se:

$$CV(\hat{\theta}(f))_{(k)} = \left(\frac{fo(1-f)}{f(1-fo)}\right)^{1/2} CV(\hat{\theta}(fo))_{(k)}, \qquad k = 1, \cdots, r$$

Desta maneira, para assegurar que um percentual mínimo αx100% das características de interesse serão estimadas com CV abaixo do limite **c** é suficiente garantir que:

$$CV(\hat{\theta}(fo))_{(t)} \leq \left(\frac{f(1-fo)}{fo(1-f)}\right)^{1/2} .C \tag{12}$$

onde

$$t = \begin{cases} r.\alpha, & \text{se } r.\alpha \epsilon N \\ [r.\alpha] + 1, & \text{se } r.\alpha \not\epsilon N \end{cases}$$

com $[r.\alpha]$ representando a parte inteira de $r.\alpha$.

Fazendo

$C_o(\alpha) = CV(\hat{\theta}(fo))_{(t)}$, obtém-se, sob AASSR:

$$\{f \mid P_c(f) \leq 1-\alpha\} = \left\{f \mid f \geq \frac{f_o C_o^2(\alpha)}{f_o C_o^2(\alpha) + C^2(1-f_o)}\right\} \tag{13}$$

Como desejamos o ínfimo deste conjunto, temos:

$$f(c,\alpha) = \frac{f_o C_o^2(\alpha)}{f_o C_o^2(\alpha) + c^2(1-f_o)} = \left(\frac{C}{C_o(\alpha)} \quad \frac{(1-f_o)}{f_o} + 1\right)^{-1} \tag{14}$$

Observe-se que

$$(i) f(0,\alpha) = 1, \quad \forall\alpha \quad \varepsilon \quad [0,1)$$

$$(ii) \lim_{c \to +\infty} f(c,\alpha) = 0, \quad \forall\alpha \quad \in \quad [0,1)$$

$$(iii) \frac{\partial f}{\partial c}(c,\alpha) = -\frac{(1-fo)}{C_o(\alpha) fo}\left(\frac{C}{C_o(\alpha}\frac{(1-fo)}{fo}+1\right)^{-2} < 0 \qquad (15)$$

Tomando-se, por exemplo, $fo = 0,25$, obtém-se

$$f(c,\alpha) = \left(\frac{3.c}{C_o(\alpha)}+1\right)^{-1}$$

Na seção seguinte são apresentados diversos gráficos resultantes da aplicação da metodologia aqui descrita a municípios de diferentes regiões brasileiras, levando-se em conta características de domicílios investigadas no Censo Demográfico de 80, a nível de município.

3. ANÁLISE E APRESENTAÇÃO DOS RESULTADOS

Nesta seção é apresentada uma aplicação da metodologia descrita anteriormente, tomando-se por base as frações amostrais que vêm sendo cogitadas para utilização no **Censo Demográfico de 1990**, a saber: 10% e 25%.

Os dados para os cálculos foram obtidos a partir das características de domicílios investigadas no Censo Demográfico de 1980 que tiveram suas estimativas divulgadas a nível de município(cerca de 65 variáveis, ao todo). Foram utilizados no estudo 48 municípios com populaç ões variando entre 1966 e 1 307 608 habitantes, escolhidos de forma a abranger todas as regiões do País.

De modo a facilitar a apresentação dos resultados, os municípios foram agrupados segundo faixas de população, apresentadas a seguir

População	Número de Municípios
< 4000	10
$4000 - 14.000$	16
$15.000 - 30.000$	10
$60.000 - 100.000$	4
$100.000 - 140.000$	4
> 400.000	4
Total	48

Conforme se observa no quadro acima, há uma predominância de municípios pequenos e médios. Esta escolha foi devida a crença, posteriormente confirmada, de que os efeitos da redução da fração amostral deveriam ser mais sentidos nos municípios com menores contingentes populacionais. Buscou-se, desta forma, analisar de maneira mais cuidadosa os impactos da redução da fração amostral nestes municípios.

Os resultados apresentados nesta seção foram inicialmente obtidos para cada município e depois agregados por faixa de população. Em nenhum momento são consideradas

agregações de faixas de população, já que, como ficará claro no transcorrer do texto, os valores obtidos guardam uma estreita relação com o tamanho da população subjacente aos cálculos.

Conforme mencionado na descrição da metodologia, as perdas de eficiência e as frações amostrais necessárias são calculadas a partir dos coeficientes de variação amostral (CVs) dos estimadores das características de interesse. Na verdade, como explicitado ao final da seção anterior, são utilizadas estimativas dos verdadeiros CVs.

Estimativas para os CVs referentes à fração de 25% encontram-se disponíveis nas publicações do Censo Demográfico de 1980, já que esta foi a fração amostral adotada naquela ocasião. A metodologia de cálculo das estimativas dos CVs dos estimadores das características investigadas no CD-80 é apresentada no anexo 3(para uma descrição mais detalhada, consultar [1]).

Os coeficientes de variação amostral correspondentes à fração de 10% são obtidos a partir dos CVs referentes à fração de 25%, através da fórmula(11) da seção anterior. Assumiu-se, neste cálculo, o esquema de $AASSR$[1].

No cálculo dos valores das perdas de eficiência, bem como das frações de amostragem necessárias, foram utilizados os seguintes limites de tolerância para o erro amostral(em %):

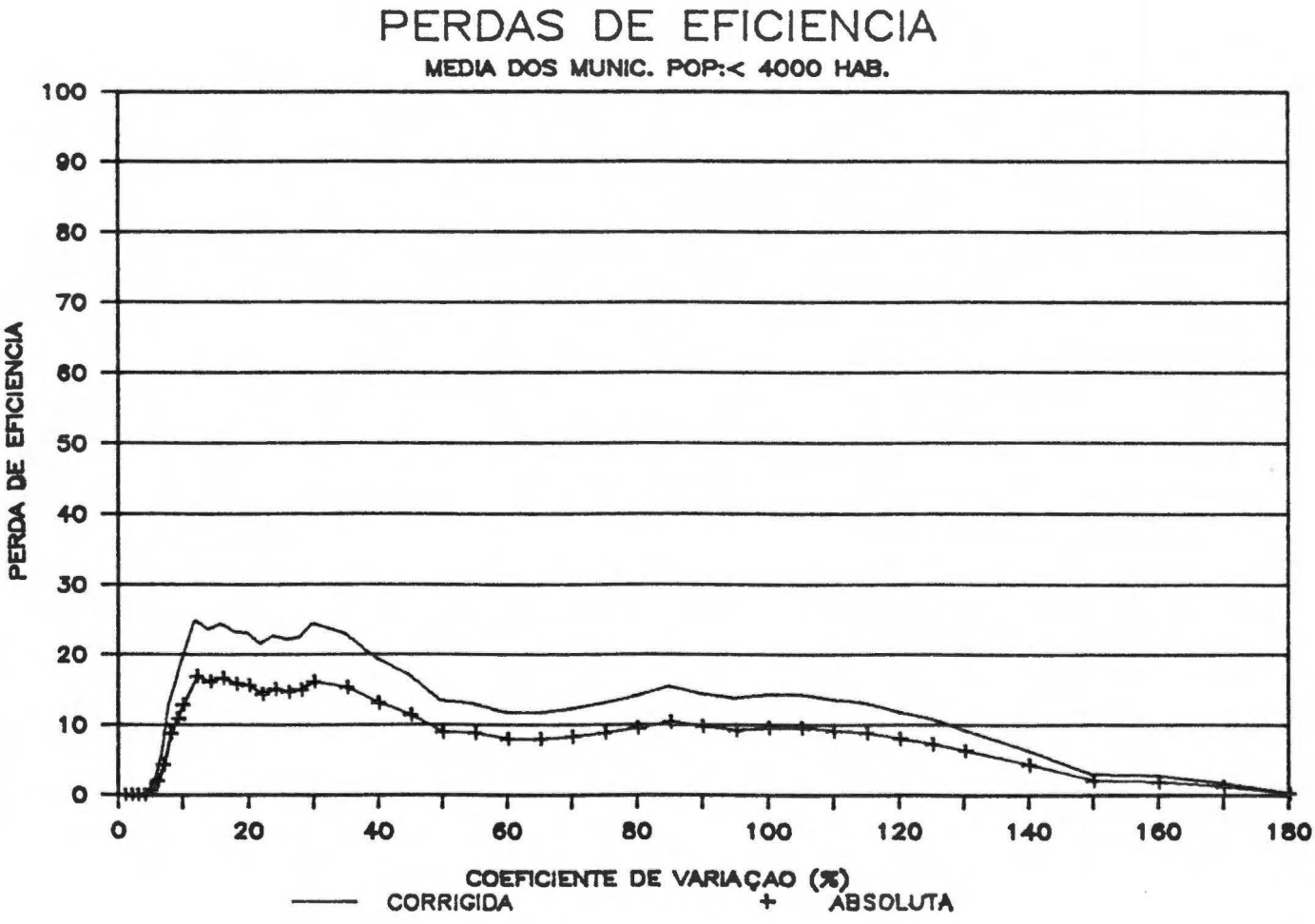
$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, \cdots, 30, 35, 40, \cdots, 130, 140, 150, 160, 170 e 180.$$

1 - Na verdade, o esquema adotado no CD-80 foi de seleção por amostragem sistemática, com expansão através de um método de pós-estratificação, denominado PIETOM (Ver [1]). Os resultados obtidos, contudo, não diferem significativamente daqueles que teriam sido encontrados, caso fosse assumido o esquema de AASSR

São apresentados a seguir, sob forma de gráficos, os principais resultados da aplicação, mantendo-se a sub-divisão por faixas de população mencionadas anteriormente. Breve comentários são incluídos junto com cada gráfico, de modo a auxiliar sua compreensão. Os comentários gerais e as análises comparativas são guardadas para a última seção.

- Municípios com população inferior a 4000 habitantes

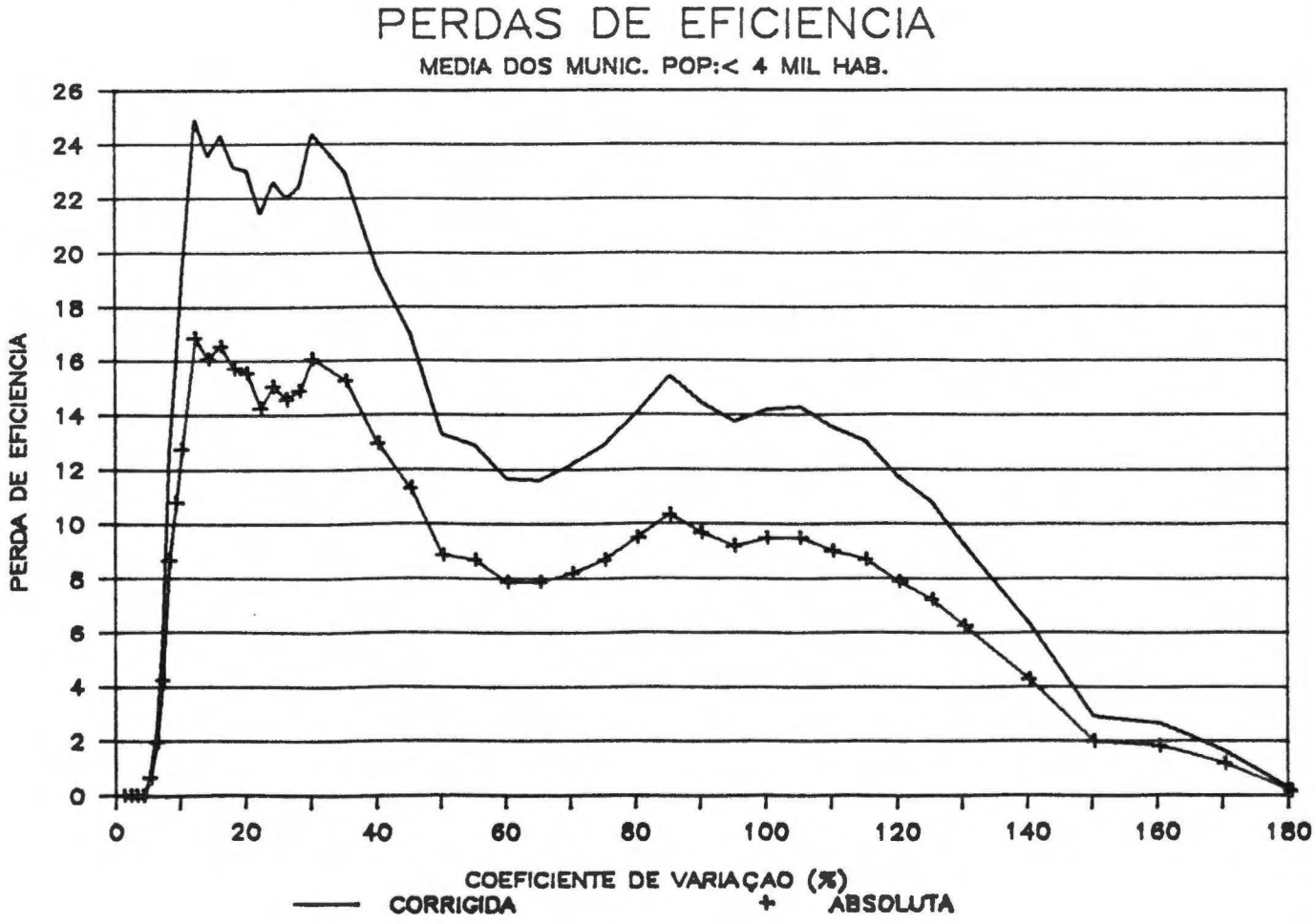
O gráfico a seguir mostra o comportamento da perda média de eficiência para as estimativas provenientes de uma amostra simulada de 10% quando comparada com os resultados obtidos com a amostra de 25% do CD-80.



Para os municípios com contingente populacional nesta faixa, a utilização de uma amos-

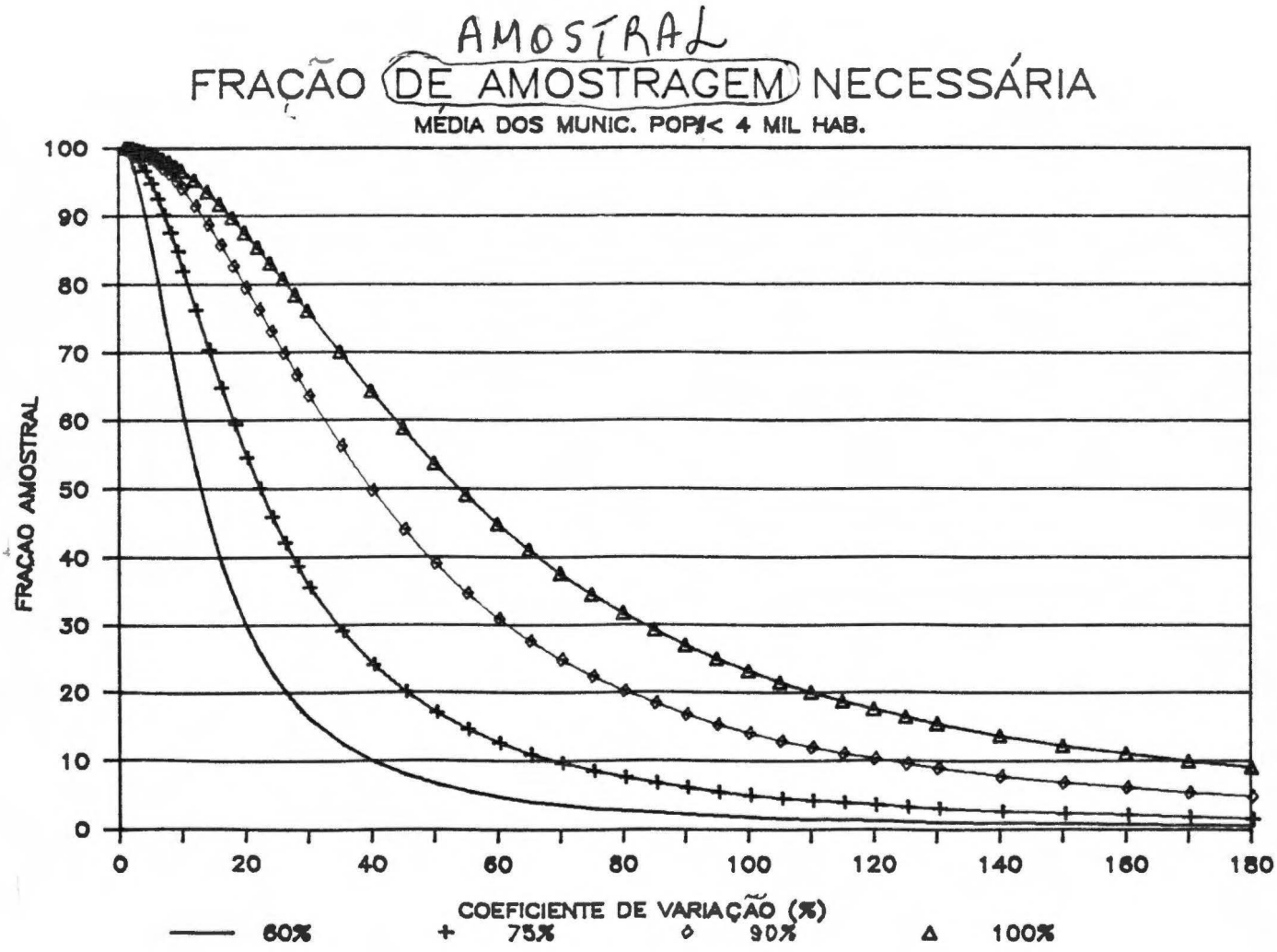
tra de 10%, em detrimento à de 25%, acarreta uma perda média com valor máximo em torno de 25% (eficiência corrigida) correspondendo às estimativas com erro amostral estimado em torno de 10%. Observa-se que para as estimativas com erro amostral estimado entre 10% e 40% correspondem as maiores perdas de eficiência, sendo que a perda de precisão para as estimativas com erro amostral superior a 40% oscila em torno de 10%, decrescendo exponencialmente para as estimativas, com erro amostral correspondente, em torno de 110%. È interessante ressaltar que mesmo para as estimativas com erro amostral estimado elevado, por exemplo, CVs superior a 40%, as maiores perdas de eficiência observadas correspondem, provavelmente. às estimativas de características que são bastante rarefeitas na população considerada. Para esses municípios existe uma diferença da ordem de, aproximadamente, 8% entre a perda média de eficiência corrigida e absoluta.

No gráfico seguinte pode-se visualisar melhor o comportamento das distribuições das perdas médias de eficiência corrigida e absoluta para esses municípios cuja população é inferior a 4000 habitantes.



Observe que a perda média máxima de eficiência corrigida é da ordem de 25% aproximadamente, enquanto que a absoluta não ultrapassa 17%. Esta diferença está relacionada com o fato de se estar trabalhando com características muito rarefeitas na população considerada.

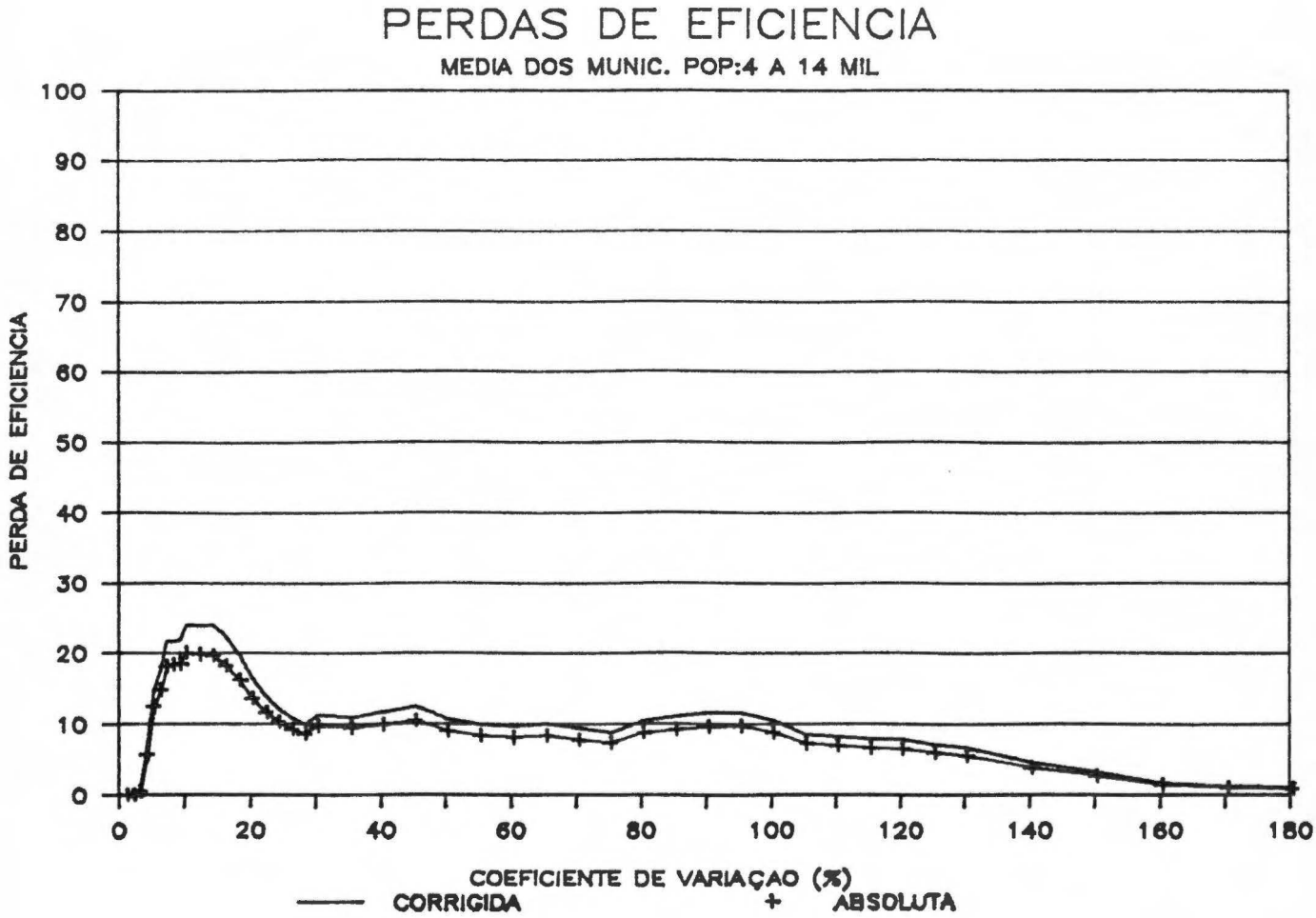
O gráfico a seguir indica a fração de amostragem necessária para se obter com um nível de precisão pré-fixado estimativas para 60%, 75%, 90% e 100% das caracterísiticas de domicílios que tiveram suas estimativas divulgadas a nível de município no Censo Demográfico de 1980. O coeficiente de variação é utilizado para fixar o nível de precisão desejado à priori.



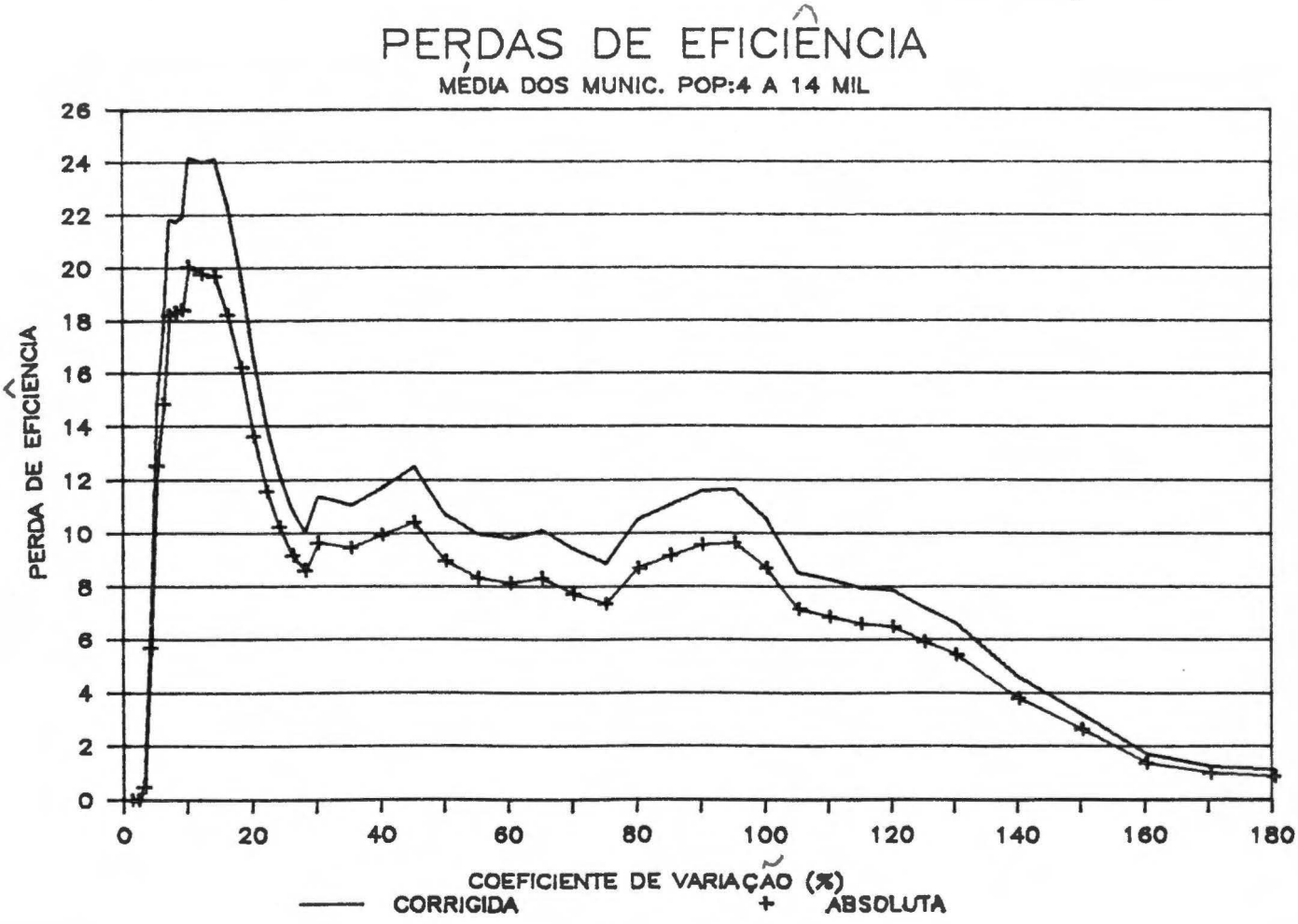
Observe que, fixado à priori um erro amostral igual ou inferior a 10%, seria necessário, em média, uma amostra de aproximadamente 60% dos domicílios do universo considerado para fornecer estimativas dentro da precisão requerida, para 60% das variáveis consideradas no estudo. Uma amostra de 30% forneceria para 75% das variáveis, estimativas com erro amostral estimado menor ou igual a 35%.

- Municípios com população entre 4000 e 14000 habitantes

A perda média máxima de eficiência das estimativas obtidas para os municípios com contingente populacional nesta faixa é de aproximadamente 25% (eficiência corrigida) e 20% (eficiência absoluta)correspondente as estimativas com erro amostral estimado entre 12 e 18% aproximadamente. Observe que para as estimativas com erro amostral estimado a partir de 18%, a perda média de eficiência (absoluta e corrigida) decresce exponencialmente.

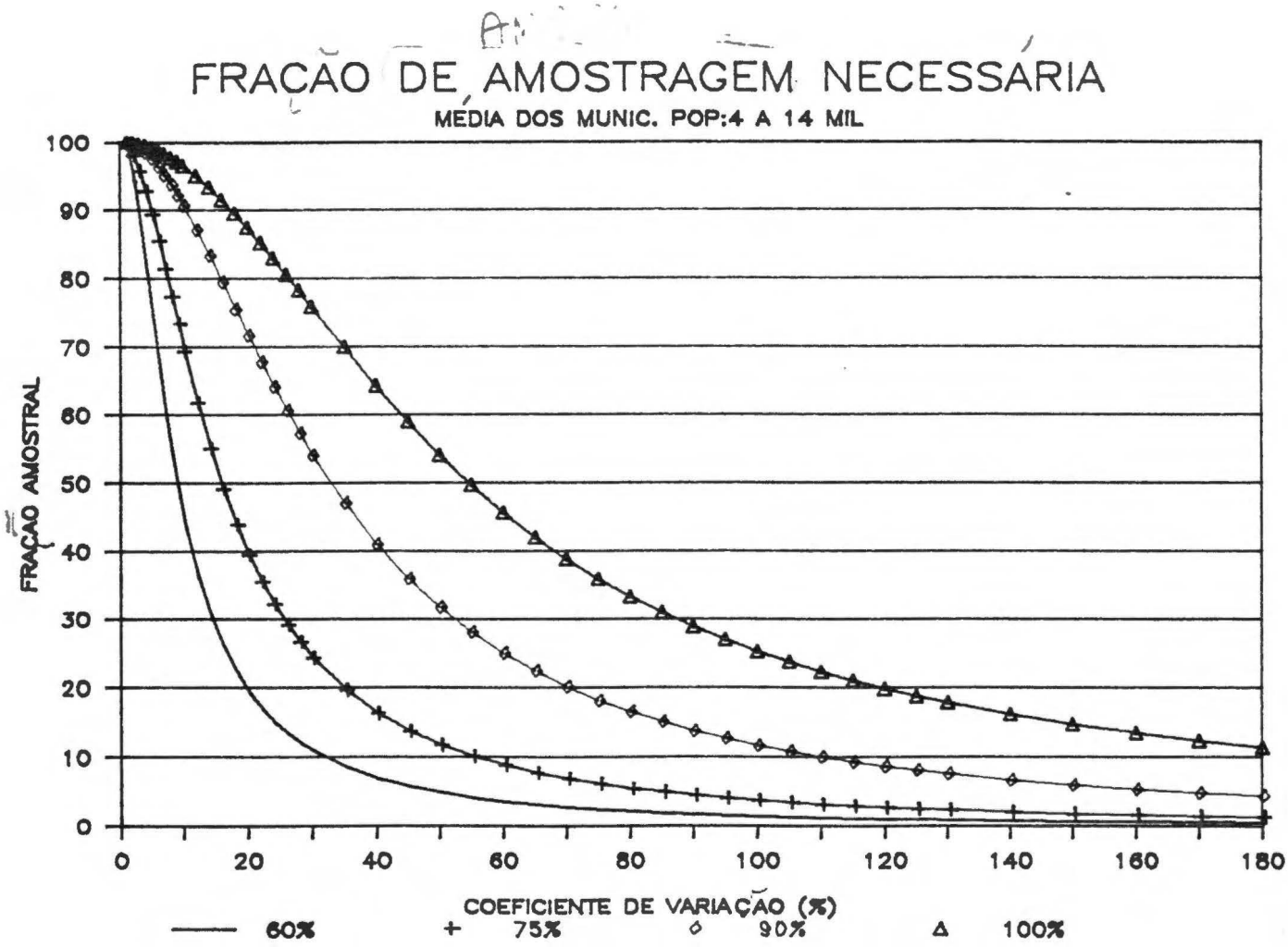


O Gráfico a seguir permite visualisar o comportamento entre as eficiências (corrigida e absoluta). A diferença entre as perdas de eficiência é da ordem de 5% e ambas apresentam comportamento exponencialmente decrescente a partir de um erro amostral estimado em torno de 18%, sendo que a perda da eficiência corrigida se estabiliza em torno de 12% para estimativas com CVs da ordem de 30%, enquanto que a absoluta se estabiliza em torno de 10%. A partir de CVs da ordem de 100% ambas as curvas decaem exponencialmente.



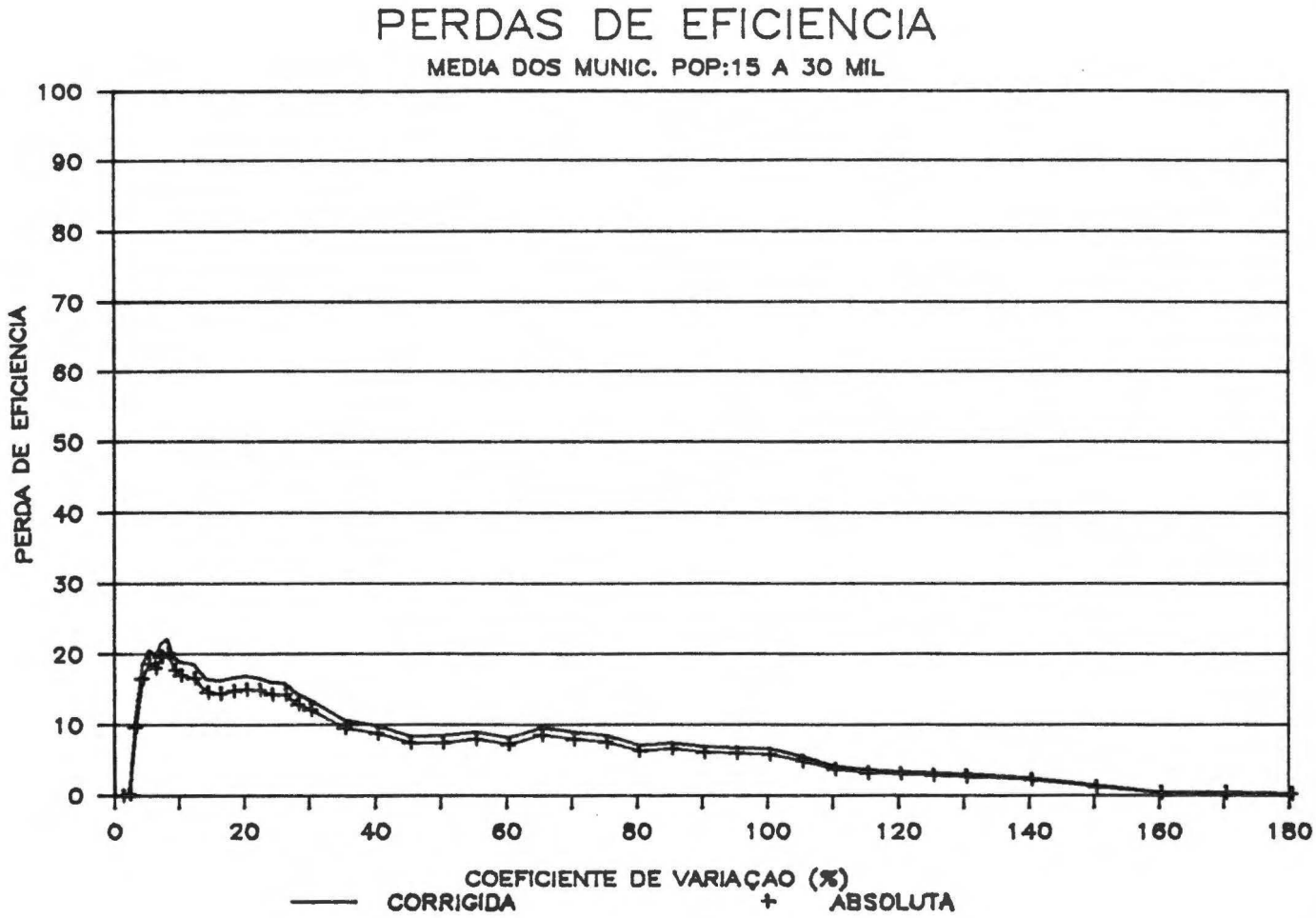
Quanto a fração amostral necessária para se obter com um nível de precisão pré-fixado estimativas para 60%, 75%, 90% e 100% das variáveis investigadas, observa-se no gráfico a seguir que se fixado um erro amostral em 10% ,seria necessário, em média, uma amostra de 50% (fração amostral igual a 1/2) para fornecer estimativas com precisão dentro da requerida para 60% das variáveis consideradas. Uma amostra de 30% forneceria em média,

para 75% das variáveis, estimativas com erro amostral estimado menor ou igual a 25%.



- Municípios com população entre 15000 a 30000 habitantes

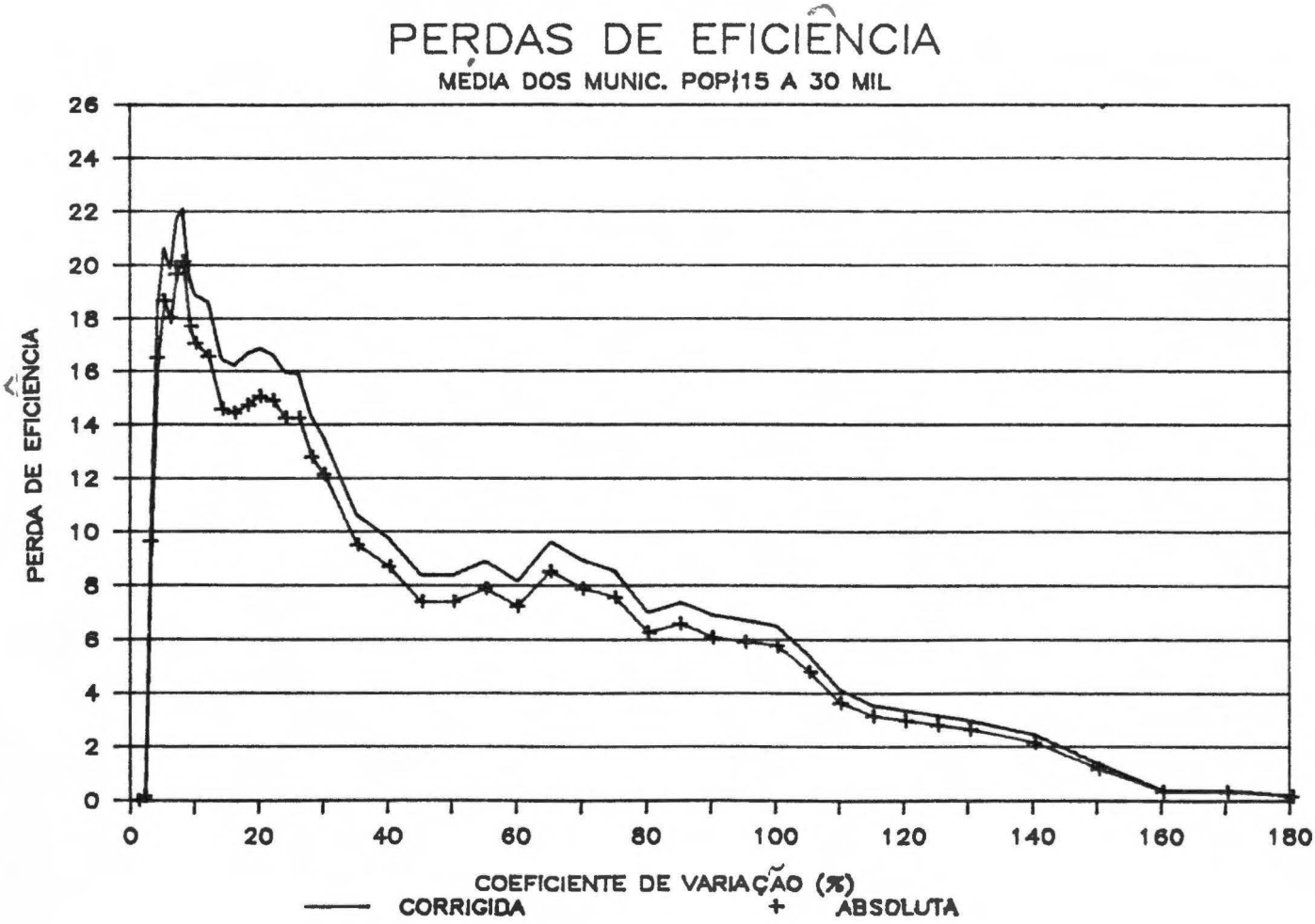
Para os municípios com contingentes populacionais nesta faixa a perda média máxima é da ordem de 22% (corrigida) e 20%(absoluta), correspondente às estimativas com erro amostral estimado em torno de 80%. Observe que para estimativas com erro amostral superior a 35%; ambas as perdas médias de eficiência são inferiores a 10%, decrescendo exponencialmente.



Observa-se, no gráfico a seguir, que a diferença entre as perdas média máxima de eficiência é da ordem de 2%, sendo que para estimativas com CVs superior a 25% as perdas de eficiência apresentam comportamento exponencialmente decrescente, permane-

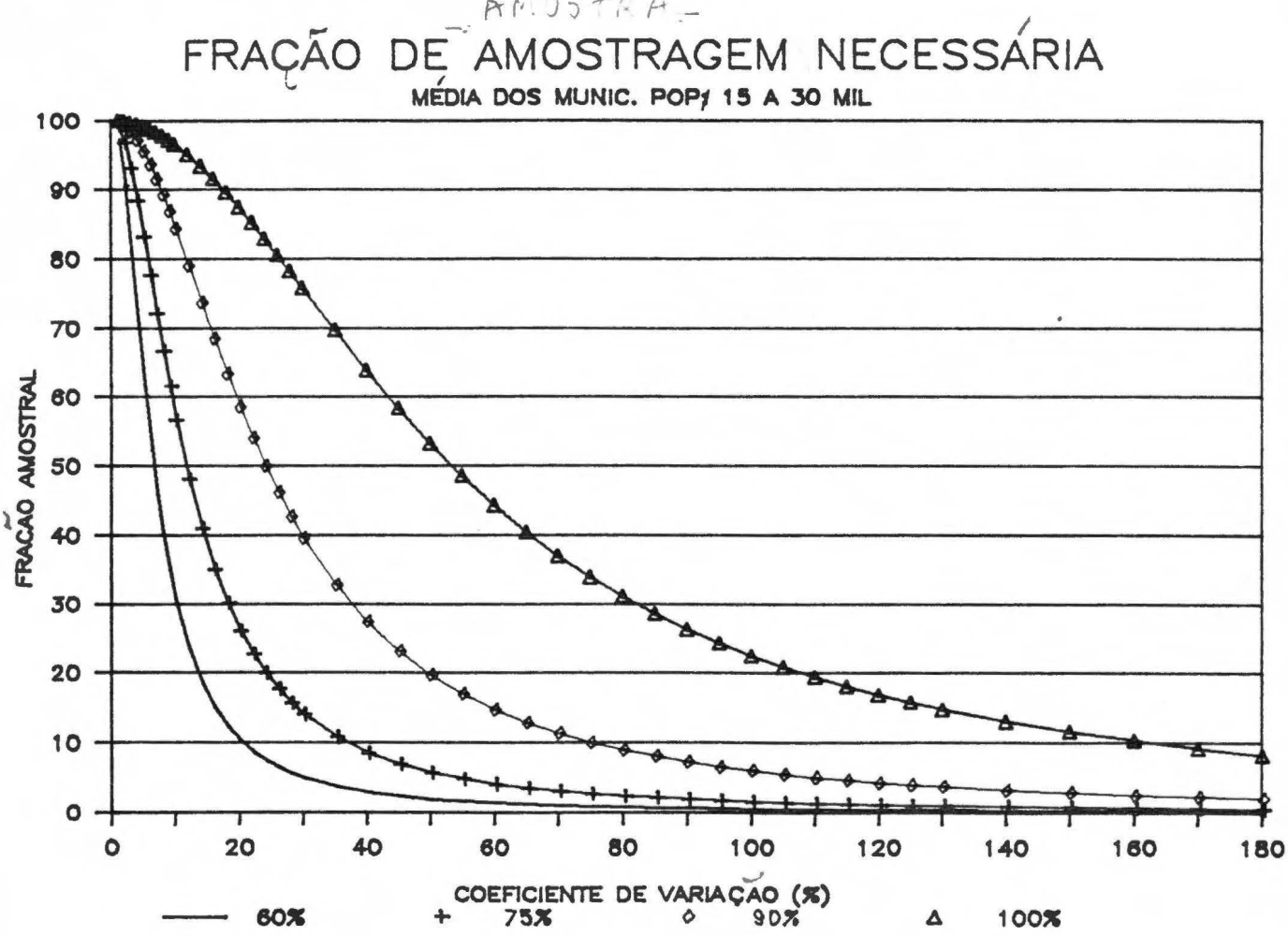
Tenha RBE

cendo estáveis em torno de 8% para aquelas estimativas com erro amostral estimado entre 45% e 75%. Note que, para estes municípios com população na faixa considerada, é pequena a diferença existente entre as perdas médias de eficiência.



Fixado um nível de precisão previamente (CV de aproximadamente 10%), observa-se no gráfico seguinte que uma amostra com fração 1/3 forneceria estimativas para 60% das variáveis investigadas com a precisão desejada. Uma amostra de 30% forneceria estimativas para 75% das variáveis com erro amostral estimado não superior a 18%.

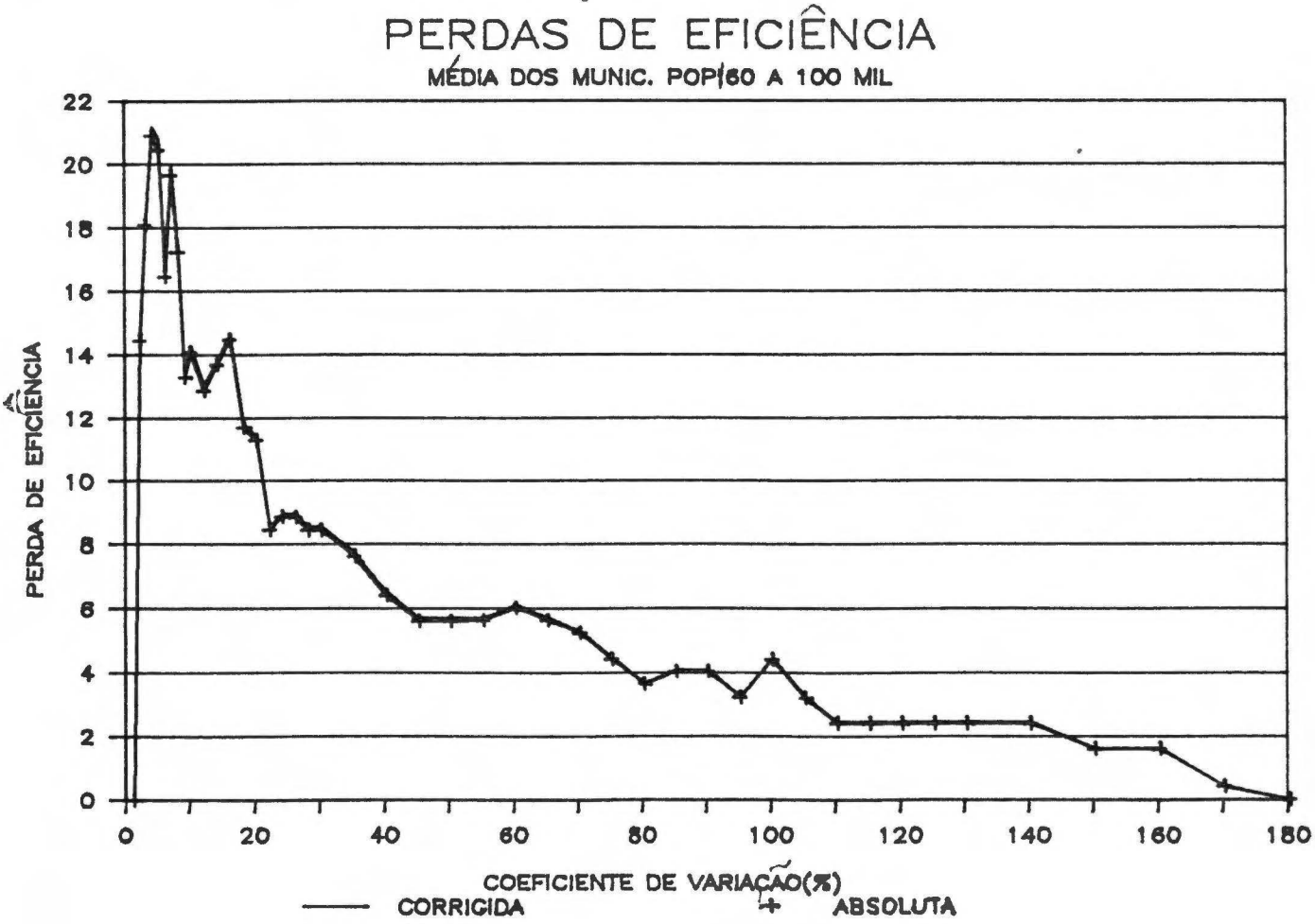
Uma amostra de 50% (f = 1/2) forneceria, em média, estimativas para 60%, 75%, 90% e 100% das variáveis investigadas, com erro amostral estimado menor ou igual a aproximadamente 8%, 13%,25% e 65% respectivamente.



- Municípios com população entre 60000 e 100000 habitantes

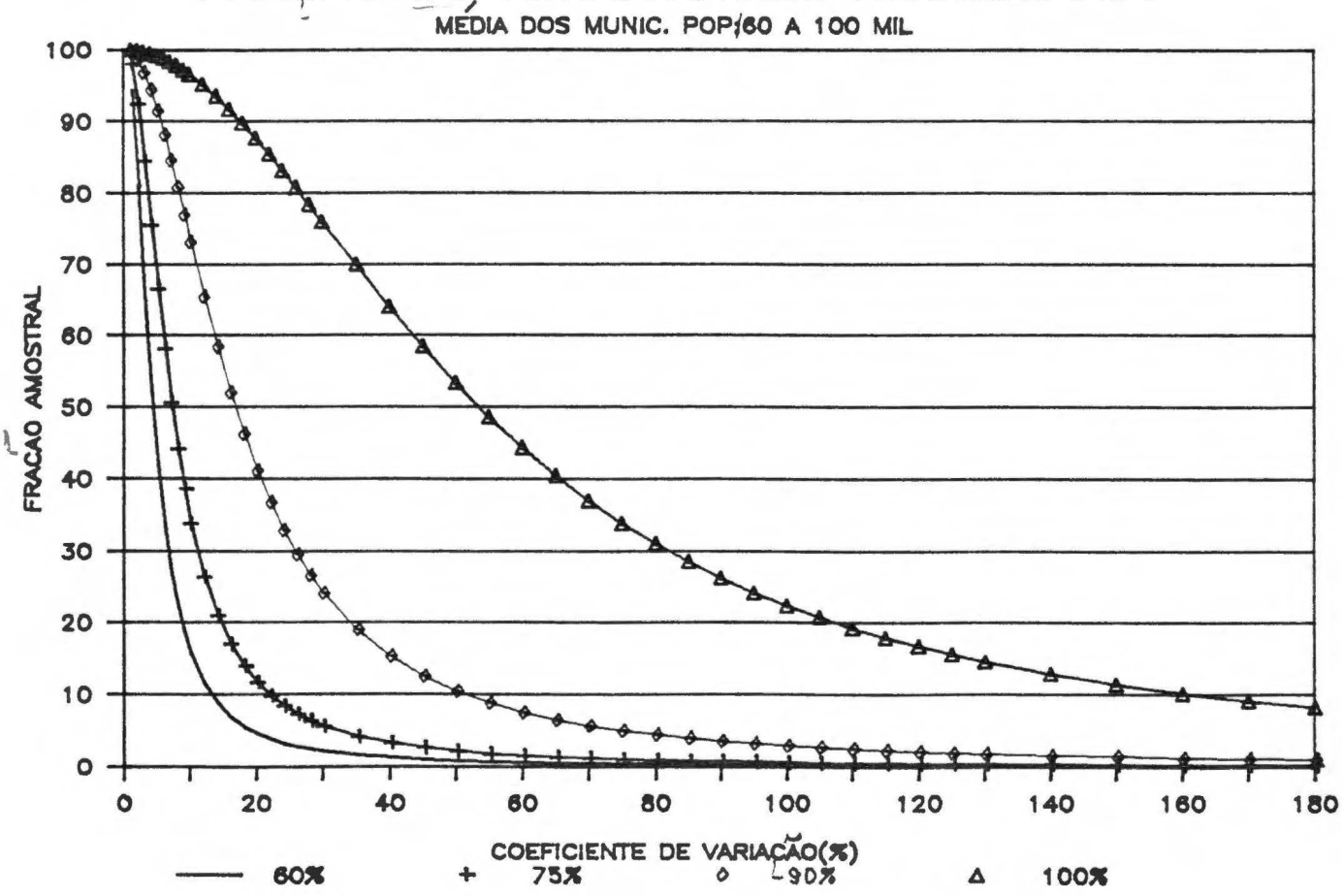
A perda média máxima de eficiência observada para os municípios com população nesta faixa não ultrapassa 21% e corresponde às estimativas com erro amostral estimado em torno de 5%. Este valor decresce à medida que se relaxa o nível de precisão, sendo que, para um erro amostral fixado previamente em torno de 20%, a perda média de eficiência já é inferior a 10%, conforme pode ser observado no gráfico a seguir.

Quanto ao comportamento das perdas de eficiência corrigida e absoluta, observa-se, no gráfico a seguir, comportamento idêntico sendo que ambas as curvas já se justapõem para os municípios com contingentes populacionais nesta faixa. A partir de um erro amostral da ordem de 15%, a perda de eficiência decresce exponencialmente já sendo inferior a 10% para estimativas com erro amostral da ordem de 20%.



No gráfico a seguir pode-se visualisar o comportamento da distribuição da fração de amostragem necessária para atender a um nível de precisão pré-determinado. Observe que as curvas correspondentes aos percentuais de variáveis consideradas no estudo apresentam comportamento acentuadamentedecrescente. As curvas decrescem exponencialmente

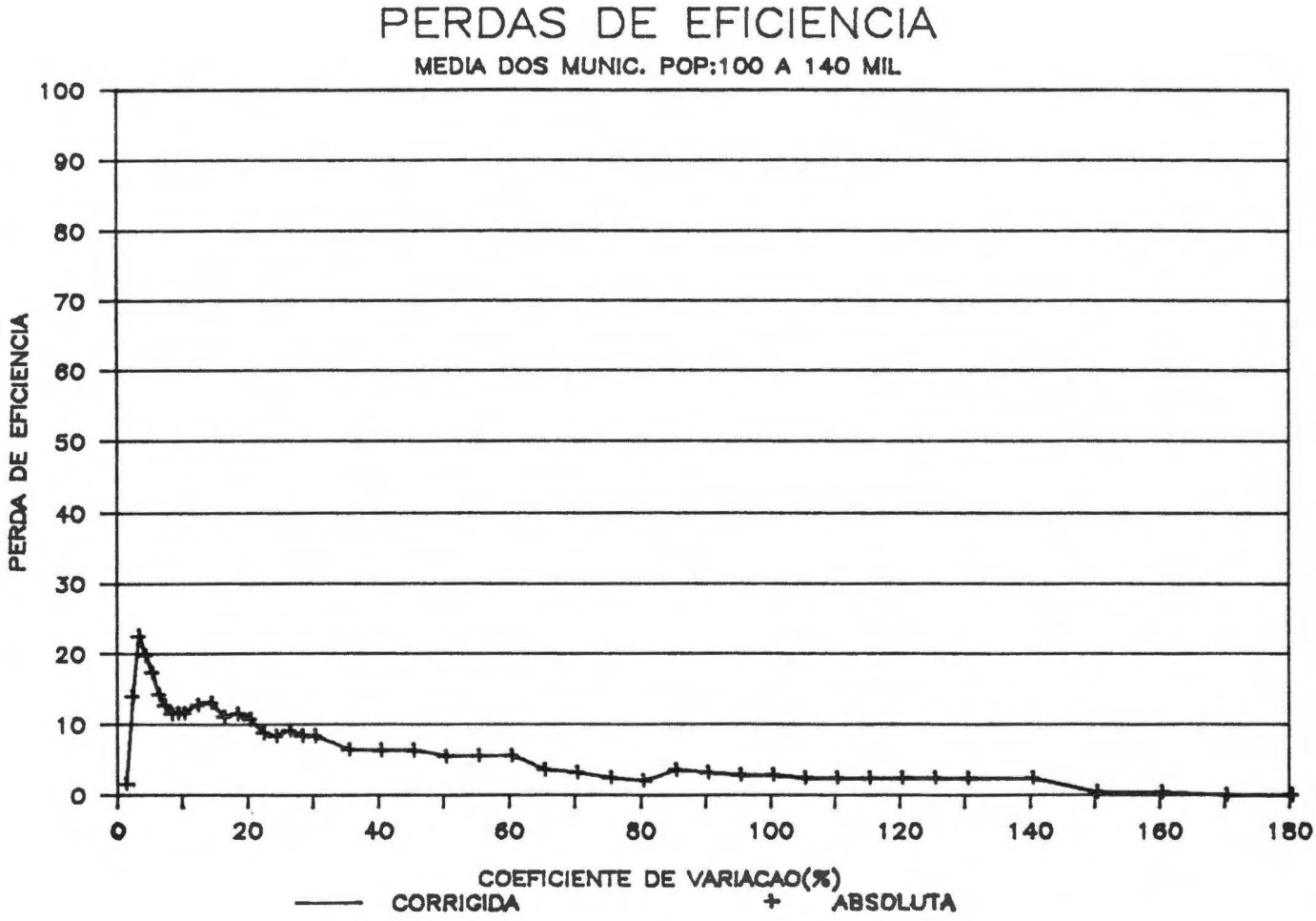
à medida que o nível de precisão é relaxado. Observe que se fixado umerro amostral em 10% seria necessário, em média, uma amostra de aproximadamente 20% (f = 1/5) para fornecer estimativas com o nível de precisão requerida para 60% das variáveis consideradas, uma amostra de 40% para estimar 75% das variáveis, e amostras de 75% e 95% para estimar, respectivamente, 90 e 100% das variáveis consideradas.



Se fixada uma fração de 10%; a amostra forneceria, em média, estimativas de 60% das variáveis com erro amostral estimado não superior a 12% e forneceria estimativas para 75% das variáveis consideradas com erro amostral estimado menor ou igual a 20%, aproximadamente.

- Municípios com população entre 100 000 e 140 000 habitantes

Para estes municípios a perda média máxima de eficiência é de aproximadamente 23% e corresponde às estimativas com erro amostral estimado inferior a 5%. A distribuição da perda média apresenta comportamento decrescente a partir de CVs da ordem de 5% permanecendo estável em torno de 10% para estimativas com erro amostral entre 10 e 20%, conforme mostra o gráfico a seguir.



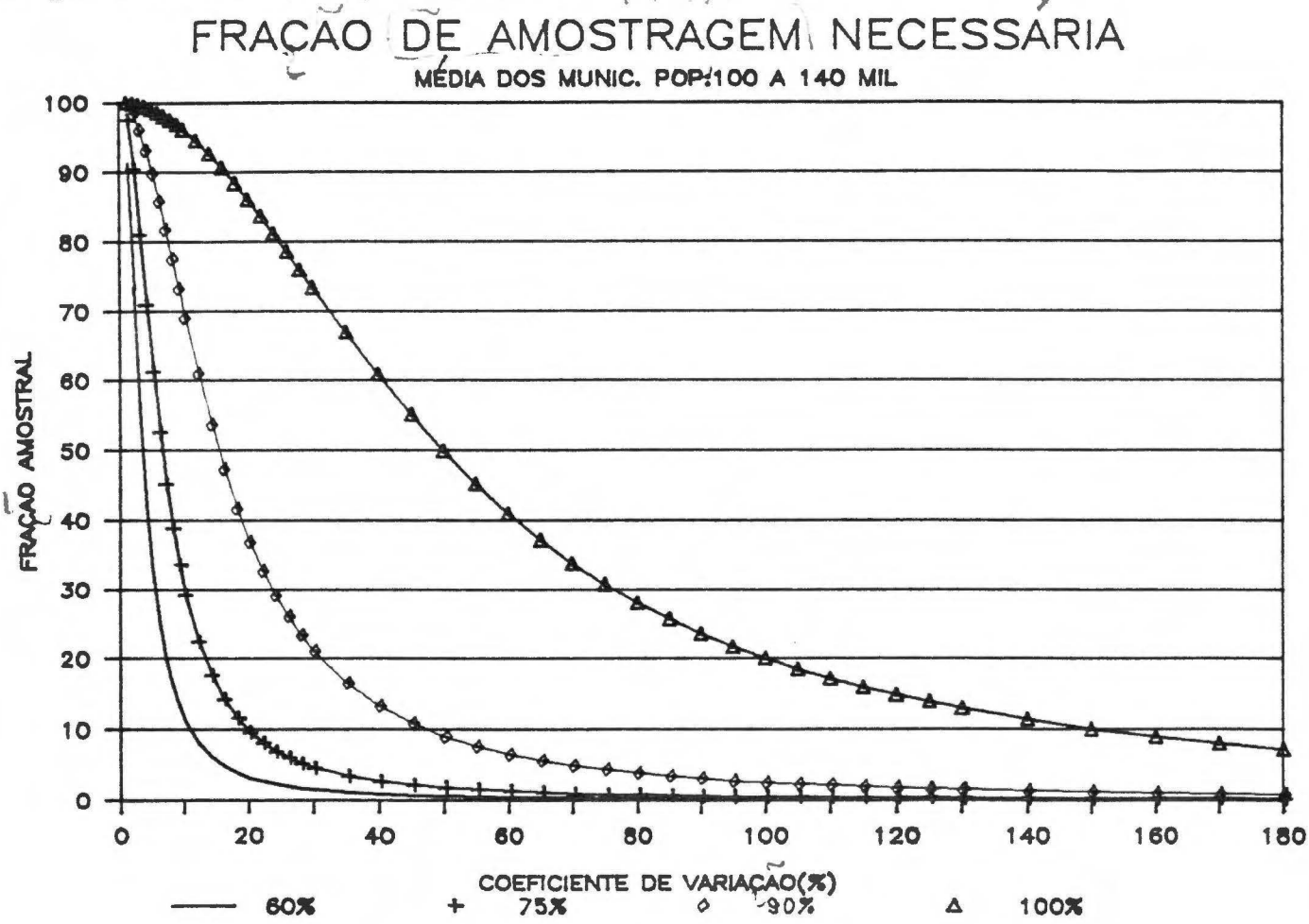
No gráfico seguinte pode-se observar a inexistência de diferença entre as perdas média de eficiências consideradas. A partir de um erro amostral estimado em torno de 20%, a perda média máxima de eficiência observada já é inferior a 10% e apresenta comportamento

Tentativa RBE

exponencialmente decrescente, vindo a permanecer estável em torno de 2% para estimativas com erro amostral estimado entre 85% e 140%.

Quanto à fração de amostragem necessária para atender um nível de precisão previamente determinado, observa-se no gráfico a seguir, um decrescimento exponencial indicando que,se fixado à priori um erro amostral em 10%, seria necessário, em média, uma amostra de aproximadamente 10% (f = 1/10) para fornecer estimativas dentro da precisão requerida para 60% das variáveis; uma amostra de 30% para estimar 75% das variáveis, e amostras de 70% e 95%, para fornecer estimativas para 90% e 100% respectivamente, das variáveis consideradas.

Observe que, se fixado uma fração amostral de 10%, a amostra forneceria em média, estimativas para 60% e 75% das variáveis consideradas, com erro amostral estimado menor ou igual a 10% e 20%, rerspectivamente.



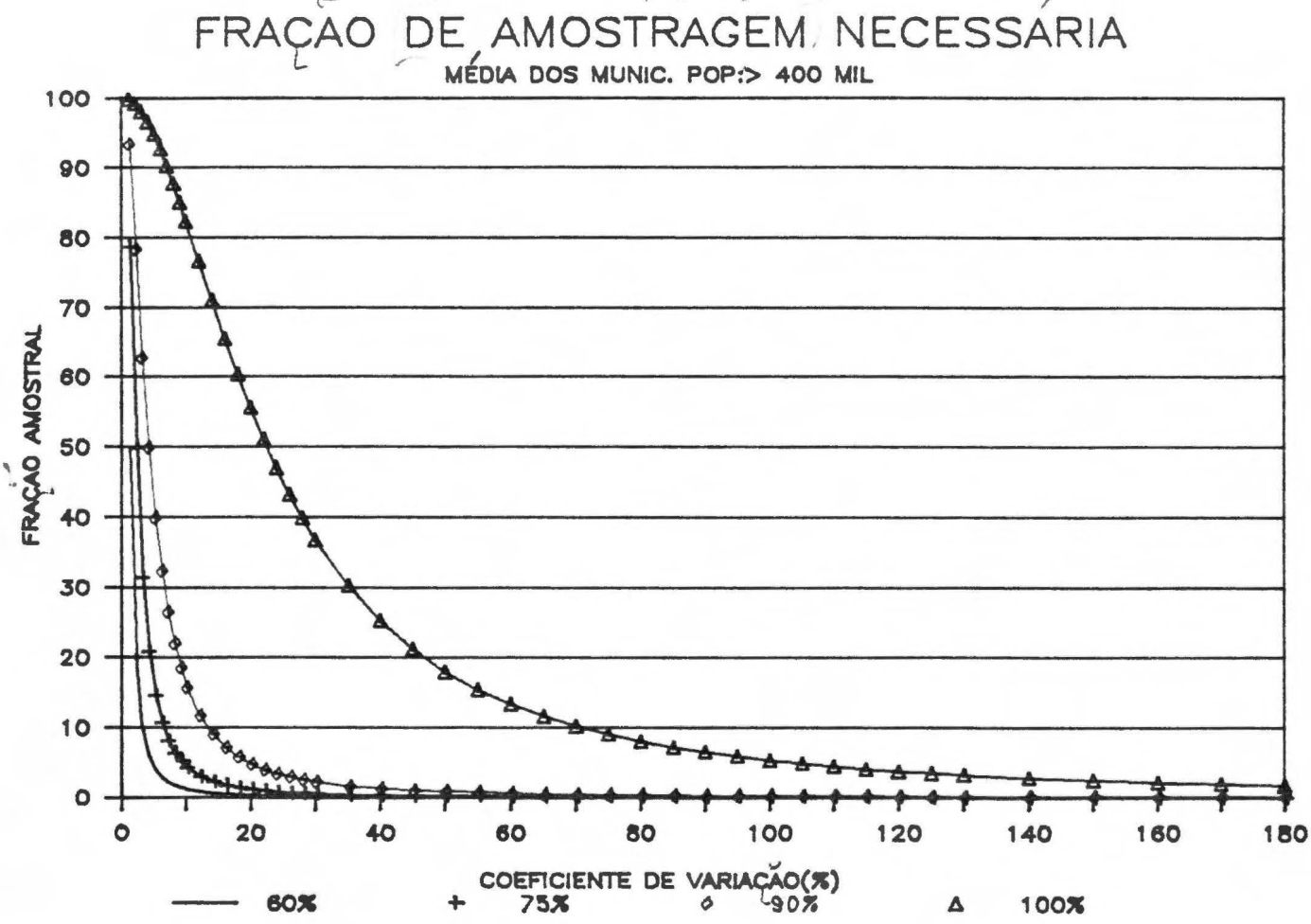
- Municípios com população superior a 400 000 habitantes

A perda média máxima de eficiência observada é de aproximadamente 24% e corresponde as estimativas com erro amostral estimado da ordem de 20%. Observa-se nos gráficos a seguir que para CVs a partir de 10%, a perda já é inferior a aproximadamente 5% e decai exponencialmente à medida que o nível de precisão é relaxado. Não existe diferença entre as perdas de eficiência considerada (as curvas se justapõem). Observe ainda que para estimativas com erro amostral estimado a partir de 30%; a perda de eficiência já é, de modo geral, aproximadamente inferior a 2%.

Quanto à fração de amostragem necessária para atender a um determinado nível de precisão fixado à priori, observa-se para os municípios considerados com contingentes populacionais desta ordem de grandeza, que o gráfico da fração de amostragem em função do erro amostral decresce exponencialmente, sendo que com exeção da curva referente à fração de amostragem necessária para estimar todo o elenco de variáveis consideradas (100%) no estudo, as demais curvas convergem para zero rapidamente.

Conforme pode-se observar no gráfico a seguir, fixado um erro amostral máximo permitido em 10%, uma amostra de aproximadamente 5% forneceria estimativas para 75% das variáveis consideradas, enquanto que para obter estimativas de 90% e 100% das variáveis, seriam necessárias amostras de 20% (f = 1/5) e 80% respectivamente.

Se fixada uma amostra com fração de 1/10, obter-se-ia, em média, estimativas para 60%, 75%, 90% e 100% das variáveis consideradas com erro amostral estimado menor ou igual a 3%,6%,12% e 70% respectivamente.



4. CONCLUSÕES

A análise das perdas de eficiência através das diferentes faixas de população consideradas no estudo indica que a perda máxima é quase independente do tamanho da população, situando-se sempre em torno de 25%. Tal perda não pode ser considerada como muito expressiva, siginificando que, na pior situação possível, um quarto das variáveis que eram bem estimadas com a fração de 25% deixaram de sê-lo com a fração de 10%.

Talvez mais importante do que a perda máxima, seja o valor do CV para o qual ela ocorre. Analisando-se esta última medida, observa-se uma forte dependência da mesma com relação ao tamanho da população. De um modo geral, a medida que aumenta a população, decresce o valor do CV associado à máxima perda de eficiência. Assim, para municípios com população até 4000 habitantes, a perda máxima ocorre para um CV de 11% enquanto que para municípios com população acima de 400000 habitantes a perda máxima se dá para CVs da ordem de 2%. Sem dúvida, isso oferece evidência de que, de um modo geral, a perda de eficiência é mais grave nos municípios pequenos do que nos grandes, o que, aliás, já era esperado.

A medida que aumenta o tamanho da população deixam de existir diferenças entre as eficiências corrigida e absoluta. Tal fato se explica em razão da menor rarefação de determinadas características em populações maiores.Esse aspecto do comportamento das características investigadas na amostra aponta para a inadequação de um plano tabular único válido para todos municípios. Aparentemente seria mais recomendável trabalhar-se com planos diferenciados,levando-se em conta a precisão com que determinadas variáveis podem ser estimadas, tomando-se por base o tamanho da população do município. Apenas variáveis que pudessem ser estimadas com um mínimo de confiança deveriam ser divulgadas.

Com relação às frações amostrais o que se oberva é um aumento considerável da velocidade de decaimento das curvas de fração de amostragem necessária com o aumento da população. A análise comparativa destas curvas leva às seguintes conclusões:

(i) - Para municípios pequenos (população menor que 14000) são necessárias frações amostrais extremamente elevadas (50% ou mais) para assegurar precisão razoável CV(20%) para estimativas de um percentual de 75% das variáveis;

(ii) - Nos municípios grandes (população maior que 400000), as frações necessárias para níveis razoáveis de precisão (CV de 10%) para a maioria das características investigadas (75%) são bem pequenas (4%);

(iii) - Para os municípios intermediários (15000 a 400000 habitantes), as frações que parecem mais apropriadas oscilam entre 10% e o valor usualmente considerado (25%).

Com base nestas observações, pode-se concluir que:

(i) - Para os municípios pequenos tanto a fração de 10% como a de 25% são inadequadas;

(ii) - Para os municípios grandes, as frações de 10% e 25% são excessivas;

(iii) - Para os municípios intermediários as frações de 10% e 25% parecem apropriadas.

Observa-se então que a adoção de uma fração de amostragem única para todos municípios não parece a decisão mais acertada (a exemplo do que já havia sido comentado com relação ao plano tabular). A utilização de frações variadas adequadas aos diferentes tamanhos populacionais, surge como uma alternativa mais atraente.

Inclui-se, a seguir, uma proposta de dimensionamento de frações de amostragem mistas, feita a partir das faixas populacionais consideradas no estudo. Tal proposta deve ser encarada mais como um exemplo da solução que poderia vir a ser adotada no censo, do que uma tentativa de estabelecer a composição ótima de frações e tamanhos populacionais.

Os objetivos principais da proposta, que doravante será denominada por **mista**, são o de

garantir maior grau de homogeneidade na precisão das estimativas, por faixa de tamanho populacional e obter redução significativa do tamanho total da amostra, com relação aquele que seria obtido com a fração de 25%.

A proposta **mista** é descrita na tabela a seguir:

Faixa de População do Município (hab.)	Fração de Amostragem
até14000	1/2
14001*a*30000	1/5
30001*a*400000	1/10
mais de400000	1/25

Em termos de precisão, a comparação entre as três propostas de dimensionamento da fração de amostragem pode ser feita com base no quadro abaixo, onde são indicados os CVs máximos esperados, por faixa de população, para 75% das características investigadas.

Faixa de População	Fração de Amostragem Proposta	Coeficiente de Variação Máximo Esperado (em %)	
do Município (hab.)	*mista*	10%	25%
até4000	22	65	40
4001a14000	16	55	30
14001a30000	23	35	20
30001a100000	21	21	12
100001a400000	20	20	11
mais de400000	10	7	4

Pode-se observar a superioridade da proposta mista nos municípios pequenos, a equivalência entre as propostas mistas e 10% nos municípios intermediários e a primazia da fração de 25% nos municípios grandes.

Verifica-se que a proposta mista mantém a precisão dentro de limites aceitáveis para todas as faixas populacionais consideradas, o que não ocorre para as duas outras propostas.

Nos municípios grandes a proposta mista perde, em termos de precisão para as frações de 10% e 25%. Entretanto, a precisão obtida com a proposta mista nestes municípios, parece ser mais do que razoável para a maioria das aplicações.

Outra comparação interessante entre as três propostas diz respeito ao custo envolvido na sua utilização. Uma análise detalhada dos custos envolvidos certamente foge aos propósitos deste estudo. Pode-se contudo, efetuar uma análise aproximada tomando-se por base o número de domicílios selecionados em cada uma das propostas. Neste ponto, abrem-se duas possibilidades: considerar-se todos os domicílios particulares ou apenas os domicílios particulares ocupados.

Na tabela abaixo são apresentados os diferentes tamanhos de amostra associados às três propostas, considerando-se o Censo Demográfico de 1980, e utilizando-se tanto os domicílios particulares em geral, como apenas os ocupados.Os valores referentes aos domicílios particulares ocupados são apresentados entre parêntesis.

Faixa de População do Município (hab.)	Tamanho da Amostra Proposta mista	10%	25%
até14000	1989545	397909	994773
	(1625519)	(325104)	(812760)
14001a30000	1025489	512745	1281861
	(850217)	(425109)	(1062771)
30001a400000	1249406	1249406	3123514
	(1065205)	(1065205)	(2663012)
mais de400000	323167	807916	2019788
	(288899)	(722246)	(1805613)
Total	4587607	2967976	7419936
	(3829840)	(2537664)	(6344156)

A partir da tabela acima pode-se observar que a proposta mista acarreta uma redução de 62% no tamanho da amostra, com relação a proposta de fração de 25%, considerando-se os domicílios particulares em geral (66% com relação aos domicílios particulares ocupados),

sendo, por outro lado, 55% maior que a amostra de 10%, relativamente aos domicílios particulares (51% com relação a domicílios particulares ocupados).

A proposta mista é uma solução intermediária entre as propostas de 10% e 25%, em termos de custo, tendo por outro lado, **um grau de homogeneidade** na precisão das estimativas das características de interesse, superior ao das duas outras propostas.

Uma possível desvantagem associada ao uso da proposta mista seria a coexistência de diferentes frações de amostragem em uma mesma pesquisa. Isso só poderia a vir a se constituir em um problema quando considerada a operacionalização da coleta no campo, já que em termos de estimação a proposta não encerra qualquer dificuldade. Contudo, tendo em vista que a fração de amostragem é única para cada município, não se acredita que a proposta mista venha trazer maiores dificuldades para coleta.

Assim sendo, a proposta mista tem como forte atrativo a garantia de precisão uniforme (dentro de limites aceitáveis) para todas as faixas de população consideradas, além de representar um ganho substancial em termos de recursos, com relação a amostra de 25%. As possíveis dificuldades operacionais que adviriam da sua utilização que, a um primeiro exame, não parecem muito siginificativas, não são sufucientes para contra-indicar sua adoção. Obviamente, a proposta mista mais adequada aos objetivos do Censo Demográfico de 1990 deve ser estudada em maior detalhe, bem como as implicações dela decorrentes. Este texto visa apenas apresentar a proposta, junto com uma análise superficial de suas vantagens, e não esgotar a discussão sobre o tema.

5. ANEXO

5. ANEXO

ANEXO 1

MUNICÍPIOS CONSIDERADOS NO ESTUDO

MUNICÍPIOS CONSIDERADOS NO ESTUDO, UNIDADE DA FEDERAÇÃO E POPULAÇÃO SEGUNDO O CENSO DEMOGRÁFICO - 80

Município considerado	UF	População
Afonso Cunha	MA	3422
Água comprida	MG	1966
Barra de São Miguel	AL	2328
Bento de Abreu	SP	2037
Caracol	MS	3819
Ilha de Paquetá(*)	RJ	2252
Mairipotaba	GO	2670
Nova Araçá	RS	2252
Vianópolis	GO	3728
Victor Graeff	RS	4125

*Região Administrativa

Abadia dos Dourados	MG	8004
Anaurilandia	MS	7222
Aparecida do Taboado	MS	14027
Bernadino de Campos	SP	8994
Cachoeira do Arari	PA	11402
Cedral	MA	12802
Dois Riachos	AL	7952
Duas Barras	RJ	7992
Monte Mor	SP	14020
Petrolina de Goiás	GO	12153
Pinheiro Machado	RS	14359
Salinópolis	PA	14349
Santa cruz do Arari	PA	4706
Santana do Mundaú	AL	13581
Sumidouro	RJ	11386
Tocantis	AM	7684

Altamira do Maranhão	MA	24722
Bujaru	PA	25992
Delmiro Gouveia	AL	26768
Guaporé	RS	24630
Mariana	MG	29401
Mirandópolis	SP	21522
Naviraí	MS	28567
Piracanjuba	GO	24095
Pirai	RJ	28789
São Mateus do Maranhão	MA	25004

Araras	SP	65010
Bacabal	MA	81361
Cachoeirinha	RS	63196
Caucaia	CE	94106

Americana	SP	121998
Bagé	RS	100133
Caxias	MA	125509
Juazeiro do Norte	CE	135620

Campinas	SP	664566
Fortaleza	CE	1307608
Porto Alegre	RS	1125478
São Luís	MA	449433

ANEXO 2

ELENCO DE VARIÁVEIS CONSIDERADAS NO ESTUDO

ELENCO DE VARIÁVEIS CONSIDERADAS NO ESTUDO

- Domicílios particulares ocupados duráveis, situação urbana
- Domicílios particulares ocupados duráveis, situação rural
- Domicílios particulares ocupados duráveis, total(urbana + rural)
- Domicílios particulares ocupados rústicos, situação urbana
- Domicílios particulares ocupados rústicos, situação rural
- Domicílios particulares ocupados rústicos, total(urbana + rural)
- Domicílios particulares permanentes já pagos
- Domicílios particulares permanentes não pagos
- Domicílios particulares permanentes alugados
- Domicílios particulares permanentes cedidos
- Domicílios particulares permanentes outra condição de ocupação
- Domicílios particulares permanentes alugados por até 1/2 salário mínimo(SM)
- Domicílios particulares permanentes com aluguel mensal maior que 1/2 SM e até 1 SM
- Domicílios particulares permanentes com aluguel mensal maior que 1 até 3 SM
- Domicílios particulares permanentes com aluguel mensal maior que 3 SM
- Domicílios particulares permanentes próprios em aquisição
- Domicílios particulares permanentes próprios com prestação mensal até 1/2 SM
- Domicílios particulares permanentes próprios em aquisição c/ prestação de 1/2 a 1 SM
- Domicílios particulares permanentes próprios em aquisição c/ prestação de 1 a 3 SM
- Domicílios particulares permanentes próprios em aquisição c/ prestação maior que 3 SM
- Domicílios particulares permanentes com canalização interna, rede geral
- Domicílios particulares permanentes com canalização interna, poço ou nascente
- Domicílios particulares permanentes com outra forma de abastecimento de água

- Domicílios particulares permanentes sem canalização interna, rede geral

- Domicílios particulares permanentes sem canalização interna, poço ou nascente

- Domicílios particulares permanentessem canalização interna, outra forma de abastecimento de água

- Domicílios particulares permanentes com abastecimento de água ignorado

- Domicílios particulares permanentes com instalação sanitária só do domicílio, rede geral

- Domicílios particulares permanentes com instalação sanitária só do domicílio, fossa séptica

- Domicílios particulares permanentes com instalação sanitária só do domicílio, fossa rudimentar

- Domicílios particulares permanentes com instalação sanitária só do domicílio, outra forma de instalação

- Domicílios particulares permanentes com instalação sanitária comum a mais de 1 domicílio, rede geral

- Domicílios particulares permanentes com instalação sanitária comum a mais de 1 domicílio, fossa séptica

- Domicílios particulares permanentes com instalação sanitária comum a mais de 1 domicílio, fossa rudimentar

- Domicílios particulares permanentes com instalação sanitária comum a mais de 1 domicílio, outra forma de instalação

- Domicílios particulares permanentes sem instalação sanitária

- Domicílios particulares permanentes com fogão

- Domicílios particulares permanentes com fogão improvisado

- Domicílios particulares permanentes com fogareiro

- Domicílios particulares permanentes que não tem equipamento para cozinha
- Domicílios particulares permanentes com equipamento ignorado
- Domicílios particulares permanentes que usam gás de botijão para cozinhar
- Domicílios particulares permanentes que usam gás canalizado para cozinhar
- Domicílios particulares permanentes que usam lenha para cozinhar
- Domicílios particulares permanentes que usam carvão para cozinhar
- Domicílios particulares permanentes que usam óleo ou querosene para cozinhar
- Domicílios particulares permanentes que usam álcool para cozinhar
- Domicílios particulares permanentes que usam eletricidade para cozinhar
- Domicílios particulares permanentes que não tem combustível para cozinhar
- Domicílios particulares permanentes com iluminação elétrica, com medidor
- Domicílios particulares permanentes com iluminação elétrica, sem medidor
- Domicílios particulares permanentes com telefone
- Domicílios particulares permanentes com rádio
- Domicílios particulares permanentes com geladeira
- Domicílios particulares permanentes com televisão a cores
- Domicílios particulares permanentes com televisão a cores e preto e branco
- Domicílios particulares permanentes com televisão preto e branco
- Domicílios particulares permanentes com automóvel para uso particular
- Domicílios particulares permanentes com automóvel para trabalho
- Domicílios particulares permanentes com tempo de residência inferior a 1 ano
- Domicílios particulares permanentes com tempo de residência no domicílio de 1 ano
- Domicílios particulares permanentes com tempo de residência no domicílio de 2 anos
- Domicílios particulares permanentes com tempo de residência no domicílio de 3 a 6 anos

- Domicílios particulares permanentes com tempo de residência no domicílio de 7 a 10 anos
- Domicílios particulares permanentes c/ tempo de residência no domicílio de 11 anos ou mais

ANEXO 3

ESTIMATIVAS DO ERRO AMOSTRAL DAS CARACTERÍSTICAS DE DOMICÍLIOS INVESTIGADAS NO CD-80

As estimativas do erro amostral referentes às características de domicílios fornecidas pela amostra do Censo Demográfico de 1980 foram obtidas, para cada área de ponderação considerada, através de um modelo de regressão ajustado a partir de um conjunto de varáveis selecionadas do plano tabular referente às características de domicílios para as quais foram calculados os erros amostrais segundo o procedimento de expansão adotado no Censo Demográfico de 80-PIETOM (ver,[1]). O anexo VIII-B em [1] apresenta as variáveis selecionadas para o cálculo dos erros amostrais, referentes às características de domicílios. Para este elenco de variáveis foram obtidas estimativas do coeficiente de variação amostral através da aplicação direta da fórmula do estimador da variância estipulada no procedimento de estimação (PIETOM) adotado.

A partir do comportamento destes erros amostrais estimados pela fórmula própria do PIETOM, foram obtidas estimativas dos parâmetros do modelo de regressão ajustado que fornecem, aproximadamente, estimativas do erro amostral associado a cada estimativa de uma dada área de ponderação (o município foi considerado área de ponderação no CD-80). Assim, para cada município, o coeficiente de variação amostral pode ser estimado, aproximadamente, através da expressão: $CV(X) = b/\sqrt{X}$; onde b é o valor do parâmetro da regressão ajustada para o município considerado e X é o valor (total) da característica de interesse. Para maiores informações, consultar [1].

6. BIBLIOGRAFIA

[1] Fundação Instituto Brasileiro de Geografia e Estatística. *Metodologia do Censo demográfico de 1980.* Série Relatórios Metodológicos/IBGE, ISSN 0101-2843, V.4 - Rio de Janeiro, 1983

[2] Costa, Luiz Nery & Lima, José Matias. *Avaliaç ão da Consistência das Estimativas de Características de Domicílios obtidas através de uma Amostra de 10% Simulada a partir da Amostra de 25% do CD-80* (Documento Interno) - DPE/DEPOP/NME,1988

[3] Bickel, Peter J, & Doksum, Kjekk A. *Mathematical Statistics: basic ideas and selected topics.* San Francisco; Düsseldorf, Holden-Day, c-1977

[4] Cochran, William Gemmell. *Sample Tecniques* - 3 rd. ed. New York J. Wiley, c-1977

NÚMEROS JÁ PUBLICADOS:

01/88 - **CRÍTICA DE RAZÕES NO CENSO ECONÔMICO**
Renato Martins Assunção (ENCE/IBGE)
Rosana de Freitas Castro (DEIND/IBGE)
José Carlos R.C.Pinheiro (ENCE/IBGE)

02/88 - **USO DE AMOSTRAGEM EM SIMULAÇÃO DE LEGISLAÇÃO TRIBUTÁRIA**
José Carlos da Rocha C.Pinheiro (ENCE/IBGE)
Manuel Martins Filho (DISUL/SERPRO)

03/88 - **FORECASTING THE NUMBER OF AIDS CASES IN BRAZIL**
Dani Gamerman (IME/UFRJ)
Helio S.Migon (ENCE/IBGE & IME/UFRJ)

04/88 - **AVALIAÇÃO DOS EFEITOS DE REDUÇÃO DA FRAÇÃO DE AMOSTRAGEM NO CENSO DEMOGRÁFICO**
José Carlos da Rocha C.Pinheiro (ENCE)
José Matias de Lima (DPE/NME)