

Pesquisa sobre Condições de Saúde Bucal da População Brasileira (SBBrasil 2003): determinação dos pesos amostrais e das informações estruturais da amostra

The Brazilian Oral Health Survey (SBBrasil 2003): determining sample weights and structural information

Rejane Christine de Sousa Queiroz ¹
Margareth Crisóstomo Portela ¹
Pedro Luis do Nascimento Silva ²
Mauricio Teixeira Leite de Vasconcellos ²

Abstract

This is an article on probability sampling written to describe the methods used to calculate and calibrate sample weights of the Brazilian Oral Health Survey (SBBrasil 2003) and identify sample structural variables. It presents an overview of the efforts made to retrieve information from the records used to select the sample units in the sampled municipalities and describes the main problems observed with the survey's database, that acted as constraints total culating the natural sample weights and identifying primary sampling units and selection strata. It compares weighted sample estimates with previously published unweighted sample descriptive statistics, concluding that the oral health of the Brazilian population was better than previously disclosed.

Oral Health; Sampling Studies; Health Status; Oral Health

¹ Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.

² Escola Nacional de Ciências Estatísticas, Fundação Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, Brasil.

Correspondência

M. T. L. Vasconcellos
Escola Nacional de Ciências Estatísticas, Fundação Instituto Brasileiro de Geografia e Estatística,
Rua André Cavalcanti 106,
Rio de Janeiro, RJ 20231-050,
Brasil.
mautlv@gmail.com

Introdução

O projeto SBBrasil 2003 foi a primeira pesquisa de abrangência nacional sobre as condições de saúde bucal da população brasileira ¹. O banco de dados encontra-se disponível no sítio do Ministério da Saúde, mas não inclui as variáveis estruturais do desenho de amostra complexo utilizado ², o que impede a produção de estimativas populacionais ponderadas, limitando seus resultados a estatísticas amostrais.

Ainda que a validade epidemiológica das estatísticas amostrais tenha sido discutida em artigo recente ³, este trabalho pressupõe que os pesos amostrais e a fonte de aleatorização devida ao desenho complexo da amostra sejam usados na obtenção de estimativas válidas para a população de pesquisa ². Os termos técnicos e conceitos básicos de amostragem que estabelecem os pressupostos deste artigo constam da Figura 1.

Buscou-se recuperar o maior número possível de informações sobre a seleção da referida amostra, visando a calcular os pesos amostrais e identificar os estratos de seleção e unidades primárias de amostragem, o conjunto de informações estruturais do desenho da amostra.

A estratégia adotada, descrita em Queiroz et al. ⁴, pressupõe o resgate das informações que deveriam constar dos relatórios das equipes de campo. Tais informações devem ser usadas em combinação com dados de escolas e creches,

Figura 1

Definições dos termos técnicos e conceitos gerais de amostragem.

POPULAÇÃO Segundo Kish²⁴, é o conceito mais básico da estatística e da pesquisa por amostragem; corresponde ao conjunto de unidades para o qual se deseja fazer inferências.

POPULAÇÃO DE PESQUISA Como definido por Cochran¹⁷, é o alvo da inferência de uma amostra probabilística. Por exemplo, a *Pesquisa Nacional por Amostra de Domicílios* (PNAD), até final da década de 1990, quando não incluía as áreas rurais da região Norte, com exceção das do estado de Tocantins, tinha como *população de pesquisa* a população brasileira menos a residente nas áreas rurais excluídas. A amostra do SBBrazil 2003 tem como *população de pesquisa* a população brasileira dos seis grupos etários incluídos na pesquisa.

AMOSTRA PROBABILÍSTICA É toda a amostra selecionada de forma que cada unidade da população tenha uma probabilidade não nula e calculável de ser incluída na amostra, denominada probabilidade de inclusão na amostra¹⁷.

AMOSTRA AUTOPONDERADA É toda amostra probabilística selecionada de forma que a probabilidade de inclusão nesta seja constante para cada unidade elementar da amostra, resultando em um peso natural do desenho constante¹⁷.

AMOSTRA COMPLEXA É qualquer amostra que inclua estratificação, conglomeração ou probabilidades desiguais de inclusão (ou combinações destes aspectos) em seu desenho^{2,13}.

PESO AMOSTRAL É o valor que permite que os dados amostrais sejam usados para inferir (ou estimar) o valor de parâmetros populacionais de interesse¹⁷.

PESO NATURAL DO DESENHO É o peso amostral calculado pelo inverso da probabilidade de inclusão na amostra¹⁷.

PESO CALIBRADO Corresponde ao peso natural do desenho ajustado por uma função de calibração, de tal modo que produza estimativas que reproduzam valores populacionais conhecidos¹⁰.

ESTIMADOR É uma função das observações amostrais usada para predizer (inferir ou estimar) um parâmetro populacional de interesse, que explícita ou implicitamente considera os pesos amostrais¹⁷.

ESTIMADOR NÃO ENVIESADO (OU NÃO TENDENCIOSO OU NÃO VICIADO) É aquele cujo valor esperado, considerando todas as possíveis amostras de mesmo tamanho e mesmo desenho, é igual ao parâmetro populacional de interesse¹⁷.

ESTIMATIVA É o valor de um estimador aplicado a uma amostra particular⁵. Diz-se que uma estimativa é *não enviesada* se o estimador usado para obtê-la for não enviesado.

ESTIMATIVAS VÁLIDAS E REPRESENTATIVAS São as obtidas pelo uso de estimadores não enviesados com dados de uma amostra probabilística¹⁷.

ESTATÍSTICA AMOSTRAL É qualquer função das observações amostrais¹⁷.

PRECISÃO DE UM ESTIMADOR É medida pela variância (desvio-padrão ou coeficiente de variação) do estimador, que é uma função da variância da característica na população de pesquisa e do tamanho e desenho da amostra. Assim, para uma dada amostra, a precisão varia de acordo com a variável considerada e o estimador usado.

INTERVALO DE CONFIANÇA Indica os limites inferior e superior de um intervalo que deve conter o valor do parâmetro populacional de interesse com probabilidade especificada. Esses limites são geralmente calculados subtraindo e somando, ao valor da estimativa pontual, múltiplos de seu erro padrão (isto é, o desvio-padrão do estimador correspondente); o coeficiente aplicado é uma abscissa da distribuição teórica do estimador determinada conforme a probabilidade de cobertura desejada para o intervalo¹⁷. Nesse sentido, apesar de haver uma só amostra, o intervalo de confiança é calculado com base na distribuição teórica do estimador para todas as possíveis amostras de mesmo desenho e tamanho.

ERRO QUADRÁTICO MÉDIO DE UM ESTIMADOR É a medida do erro total que corresponde à soma da variância com o quadrado do vício, ou seja, uma função conjunta dos erros de medida, dos vícios introduzidos no processo de amostragem e da variância amostral¹⁷.

A teoria clássica de amostragem probabilística mostra que estimadores que consideram os pesos amostrais são não enviesados para estimar totais, médias, proporções e razões populacionais^{10,17}. Por exemplo, médias amostrais ponderadas são não enviesadas para estimar a média populacional correspondente. Em contraste, em amostras não autoponderadas (como é o caso da usada no SBBrazil 2003), simples estatísticas descritivas amostrais (como a média amostral ou média aritmética simples das observações amostrais) são enviesadas para estimar a média populacional.

Essa mesma teoria estabelece distribuição assintótica normal de estimadores que consideram os pesos amostrais, o que permite calcular medidas de precisão das estimativas e intervalo de confiança para os correspondentes parâmetros populacionais.

Propriedades semelhantes não existem para estatísticas descritivas amostrais que ignoram pesos ou outras informações estruturais do desenho amostral empregado para obter a amostra.

Como tais estatísticas amostrais são enviesadas e não se conhece sua distribuição de probabilidades, elas não fornecem estimativas válidas para parâmetros populacionais, e sua interpretação deve ser restrita à análise descritiva da amostra realizada.

coletados no *Censo Escolar 2003*, realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), do Ministério da Educação, bem como de setores, quadras e domicílios do *Censo Demográfico 2000*, conduzido

pela Fundação Instituto Brasileiro de Geografia e Estatística (IBGE).

Trata-se, portanto, de um artigo de amostragem probabilística, que descreve os métodos para determinação das variáveis estruturais da

amostra e apresenta a forma indicada de estimação com base nesses dados.

Métodos

Recuperação das informações sobre a seleção da amostra

A recuperação foi realizada nos municípios amostrados e dirigida no sentido de obter o máximo de informação possível sobre número de escolas, creches e alunos existentes no cadastro de seleção utilizado; relação das escolas e creches de fato selecionadas e setores censitários realmente visitados na pesquisa (os códigos na base de dados do SBBrazil 2003 não coincidem totalmente com os setores selecionados pela equipe central); número de quadras/vilas rurais consideradas na seleção em cada setor ou município com população de até 50 mil habitantes em 1º de julho de 1999; número de domicílios selecionados; e período de coleta de dados, já que a pesquisa ocorreu entre 2002 e 2003. Assim, adotou-se uma estratégia de busca dos relatórios municipais da pesquisa a partir do contato com a cadeia hierárquica desta, começando pela Coordenação Nacional de Saúde Bucal, Ministério da Saúde (COSAB/MS), passando pelos coordenadores regionais, estaduais e municipais, até o contato com participantes que coletaram dados.

Verificado que não seria possível obter os dados de todos os 250 municípios da amostra, decidiu-se priorizar o resgate das informações sobre as capitais, em particular os dados de identificação dos estabelecimentos de ensino e setores censitários efetivamente pesquisados. Essa decisão justificou-se porque os municípios de capital, por terem sido incluídos na amostra com certeza, são estratos de seleção, e suas escolas, creches e setores censitários são unidades primárias de amostragem; por outro lado, os demais municípios são unidades primárias de amostragem, bastando conhecer o número de unidades na população e na amostra para calcular os pesos amostrais.

Crítica das variáveis de identificação do banco de dados

Para gravar as informações estruturais do desenho da amostra no banco de dados do SBBrazil 2003, foi necessário assegurar que este não tivesse problemas em suas variáveis de identificação. Para tanto, adotaram-se procedimentos de crítica usuais em pesquisas domiciliares: verificação de valores válidos, testes de sequência e registros duplicados^{5,6}. Os programas foram desenvolvi-

dos usando o SAS versão 9.1.3 (SAS Inst., Cary, Estados Unidos), licenciado para a Escola Nacional de Ciências Estatísticas, IBGE.

Cálculo dos pesos amostrais

As probabilidades de inclusão na amostra, descritas em Queiroz et al.⁴, derivam das contagens de unidades na população e na amostra. Sempre que as informações recuperadas indicavam os quantitativos usados na seleção da amostra, as probabilidades de inclusão foram calculadas com esses valores. Na sua ausência, os dados populacionais foram obtidos no *Censo Escolar 2003* ou na Base Operacional Geográfica (BOG) do *Censo Demográfico 2000*⁷, enquanto os dados amostrais foram contados no banco de dados do SBBrazil 2003, seguindo as regras indicadas neste artigo. Calculadas as probabilidades de inclusão, os pesos naturais do desenho foram determinados pelo inverso do produto das probabilidades de inclusão em cada estágio de seleção.

Cálculo dos pesos amostrais em domicílios

O número de domicílios por quadras ou vilas rurais da amostra foi obtido por aproximação, tendo em vista que a residência não foi uma unidade identificada na ficha de exame e, portanto, no banco de dados.

Por meio das variáveis socioeconômicas (número de pessoas no domicílio, renda familiar, posse de automóvel e número de cômodos, principalmente), utilizadas para os grupos etários a partir dos 15 anos, foram marcados conjuntos de registros com a mesma informação socioeconômica. Cada conjunto foi considerado como um domicílio, cuja contagem foi, então, feita por grupo etário; o total de domicílios na amostra, contado por quadras ou vilas rurais, foi o maior valor dentre os obtidos para os grupos etários considerados (15-19 anos, 35-44 anos e 65-74 anos). Assim, a probabilidade de inclusão do domicílio na quadras ou vilas rurais foi obtida dividindo, conforme uma ou outra, o número obtido na contagem de domicílios da amostra pelo existente na BOG do *Censo Demográfico 2000*. Nos casos em que o código de quadras ou vilas rurais era inválido, foi usada, como probabilidade de inclusão dos domicílios nestas, a razão entre o total de domicílios contados na amostra (para o conjunto de quadras ou vilas rurais não identificadas) e o total correspondente da população obtido na BOG do *Censo Demográfico 2000*.

A probabilidade de inclusão da quadras ou vilas rurais (no setor ou no município) foi obtida pela razão entre o número de quadras ou vilas rurais na amostra e o número correspondente

na população. Note que, para municípios de até 50 mil habitantes, não houve seleção de setores, portanto o número de quadras ou vilas rurais na população correspondia ao total de quadras ou vilas rurais do município; por outro lado, para os demais municípios, o valor populacional era contado no setor censitário selecionado. O valor resultante dessa razão foi aplicado a todas as quadras ou vilas rurais, já que sua seleção foi equiprovável. Para os registros com o código de quadras ou vilas rurais inválido, considerou-se, como tamanho da amostra efetiva, o tamanho previsto da amostra de quadras ou vilas rurais (segundo as regras do Manual do Coordenador⁸) e o total de quadras ou vilas rurais existente na BOG do *Censo Demográfico 2000*, de acordo com o porte populacional do município: o total de quadras ou vilas rurais do município ou o de quadras ou vilas rurais no setor.

A probabilidade de inclusão dos setores no município foi obtida de forma análoga (setores na amostra/setores na BOG do *Censo Demográfico 2000*), contando-se os setores na amostra (os identificados) ou usando como *proxy* o número daqueles selecionados pela equipe central. Para o número de setores na população, foram contados os da BOG do *Censo Demográfico 2000* que tinham domicílios, excluindo-se os especiais, como de quartéis, indústria, hospitais etc. A probabilidade de inclusão dos municípios na amostra já havia sido calculada⁴.

Desse modo, usando os dados exatos (obtidos nas informações recuperadas) ou os aproximados pelas regras descritas anteriormente, a probabilidade de inclusão de cada pessoa selecionada em domicílio foi calculada pelo produto das probabilidades de inclusão: do município, da quadra ou vilas rurais no município e do domicílio na quadra ou vilas rurais, se o município tivesse até 50 mil habitantes; do município, do setor no município, da quadra ou vilas rurais no setor e do domicílio na quadra ou vilas rurais, para municípios com mais de 50 mil habitantes. Os pesos naturais do desenho da pessoa correspondem ao inverso da sua probabilidade de inclusão na amostra.

Cálculo dos pesos amostrais em estabelecimentos de ensino

A probabilidade de inclusão dos alunos em municípios com até 20 escolas foi igual à razão entre os números de alunos de 12 anos do município na amostra e no *Censo Escolar 2003* (alunos matriculados na escola e nascidos em 1991), visto que não houve seleção de escola. A probabilidade de inclusão do aluno corresponde, nesse caso, ao produto da probabilidade de inclusão do

município pela do aluno no município. A mesma regra foi aplicada aos alunos de creches nos municípios com até 20 creches (alunos nascidos em 1998 no *Censo Escolar 2003*).

Para cidades com mais de 20 instituições de ensino, simulou-se, no *Censo Escolar 2003*, a seleção de 20 escolas com probabilidade proporcional ao número de alunos de cada uma, respeitando-se a proporção de instituições públicas e privadas, observada na amostra (variável escola). Para essas escolas, foi feita a alocação do tamanho da amostra efetiva, por instituição, de forma proporcional ao número de alunos de 12 anos em cada uma, já que a regra de seleção⁸ estabelecia que se fizesse seleção sistemática no conjunto de alunos de 12 anos de todas as escolas eleitas. Calculou-se, para cada município, a probabilidade de inclusão do aluno na escola pela razão entre o número de alunos na amostra e o número de alunos de 12 anos no conjunto de instituições escolhidas. A probabilidade de inclusão final do aluno corresponde ao produto das probabilidades de inclusão do município, da escola no município e do aluno na escola. Os mesmos métodos foram aplicados aos alunos selecionados em creches nos municípios com mais de 20 creches.

Para esse cálculo, foi necessário associar os registros dos alunos aos das escolas/creches. Utilizou-se o algoritmo de Hájek⁹, de forma que os registros de alunos foram associados às escolas/creches, respeitando-se a ordem crescente dos números aleatórios gerados, a ordem das escolas/creches na lista de seleção gerada e o tamanho da amostra na escola/creche.

Para municípios de capital, a associação entre alunos e escolas/creches foi utilizada também para definição do código das unidades primárias de amostragem. Resultados de simulação estatística feita com os dados da própria pesquisa, para diferentes associações aleatórias entre registros de alunos e escolas/creches, indicam que essa associação não tem influência relevante nas estimativas das variâncias. No entanto, informações parciais recuperadas alteraram o procedimento descrito: quando o número de escolas/creches era conhecido, ele foi usado no lugar de 20 (número previsto nas regras de seleção da amostra); quando a relação de escolas/creches selecionadas foi recuperada, o procedimento foi aplicado para as escolas/creches efetivamente escolhidas para a amostra.

Assim, usando os dados exatos (informações recuperadas) ou aproximados (regras descritas acima), os pesos naturais do desenho do aluno correspondem ao inverso da sua probabilidade de inclusão na amostra.

Cálculo dos pesos amostrais para situações especiais

As informações recuperadas indicaram regras de seleção diferentes das previstas⁸, impondo ajustes nos métodos descritos, a fim de calcular probabilidade de inclusão mais próxima da realidade de seleção.

Para crianças (18-36 meses) selecionadas na fila de vacinação (um município), pressupôs-se que todas estas residiam no município e que todos os residentes do grupo etário foram para a fila de vacinação. Com esses pressupostos, a probabilidade de inclusão foi igual à razão entre os números de crianças (18-36 meses) na amostra e na estimativa de população do município. Para crianças de cinco anos selecionadas em creches em seis municípios com até 50 mil habitantes, procedeu-se de forma análoga à descrita para a seleção de crianças em creches de municípios com mais de 50 mil habitantes. Para pessoas de 15-19 anos selecionadas em escolas (um município), adotou-se o procedimento usado para crianças de 12 anos, mantendo-se as mesmas instituições consideradas na seleção de crianças de 12 anos do município. Para crianças de cinco anos selecionadas em domicílio em sete municípios com mais de 50 mil habitantes, foram aplicadas as regras de cálculo de probabilidade de inclusão e peso usadas nesse tipo de seleção.

Para seleção baseada em cadastro de ruas (dois municípios), foram usadas as folhas de coleta do *Censo Demográfico 2000*, que indicam os nomes dos logradouros, para replicar o processo de seleção e calcular as probabilidades de inclusão na amostra. No caso da seleção a partir das listas do Programa de Saúde da Família (três municípios), a probabilidade de inclusão foi estabelecida, por município e grupo etário, pela razão entre o número de pessoas na amostra e na estimativa de população do município.

Nos casos de ampliação do número de quadras ou vilas rurais para alcançar o tamanho de amostra mínima (dez municípios) e de exclusão de escolas selecionadas por inviabilidade de deslocamento ou pelas férias escolares (dois municípios), o cálculo das probabilidades de inclusão tomou por base os quantitativos informados pelos municípios.

Identificação dos estratos de seleção e das unidades primárias de amostragem

Na descrição do desenho de amostra¹, consta que foram definidos 25 estratos, combinando cinco macrorregiões com cinco faixas de porte populacional dos municípios, de modo a asse-

gurar que as capitais das Unidades da Federação (UF) fossem incluídas com certeza na amostra. Essa inclusão com certeza das 27 capitais tem duas consequências: a primeira é o fato de que os 27 municípios passaram a ser estratos de seleção; a segunda é que, no Nordeste, onde há nove capitais, um único município foi selecionado no estrato de seleção de maior porte, tornando necessário juntar todos os seus municípios com os do estrato de seleção de porte populacional inferior e corrigir o tamanho de amostra de 10 para 11 municípios no estrato de seleção resultante: 50.001 habitantes ou mais.

Portanto, a amostra tem, de fato, 51 estratos de seleção: 27 municípios de capital e 24 estratos definidos pela combinação de macrorregião com os grupos de porte populacional⁴. Nesses 24 estratos, cada um dos 223 municípios selecionados corresponde a uma unidade primária de amostragem. Nos 27 municípios de capital, as unidades primárias de amostragem são 139 setores, 399 creches e 438 escolas da amostra.

Calibração dos pesos amostrais

Objetivando minimizar os efeitos das hipóteses adotadas, nos casos em que foi necessário usar aproximações para calcular a probabilidade de inclusão e respectivo peso, decidiu-se calibrar os pesos naturais do desenho a fim de assegurar a coerência das estimativas do SBBrazil 2003 com os totais conhecidos da população de pesquisa, por sexo e grupo etário para cada estrato de seleção. Justificativas e métodos associados à calibração de pesos amostrais já foram descritos e aplicados^{10,11,12}.

A idéia básica da calibração é estimar (por regressão) fatores de calibração, os quais multiplicam os pesos naturais do desenho para determinar pesos calibrados que tenham a propriedade de minimizar as diferenças entre as estimativas que produzem e os totais populacionais conhecidos para um conjunto de variáveis auxiliares de calibração. Neste estudo, optou-se pela aplicação de um caso particular do estimador de regressão, denominado estimador de pós-estratificação^{13,14}; nele, os fatores de calibração correspondem às razões – calculadas por estrato de seleção, sexo e grupo etário (variáveis auxiliares da calibração) – entre o total populacional conhecido e a estimativa produzida pelos pesos naturais do desenho.

O total populacional corresponde à estimativa de população para 1º de julho de 2003, distribuindo a população total dos estratos de seleção por sexo e grupo etário, empregando a mesma técnica aplicada pelo IBGE¹⁵ para distribuir a população brasileira (estimada pelo

método das componentes) entre as UF e entre os municípios.

Obtenção de estimativas

A amostra do SBBrazil 2003 é complexa, pois utiliza estratificação e conglomeração. Isso obriga que em sua análise seja utilizada a fonte de aleatorização advinda do desenho complexo da amostra^{2,13,16}. Nesse tipo de desenho, estimativas pontuais não enviesadas de totais, médias, razões, proporções e percentis podem ser obtidas apenas com o emprego dos pesos amostrais^{2,13,16,17}. Para tanto, basta usar as rotinas que aceitam ponderação de qualquer sistema estatístico, indicando a variável que contém o peso natural do desenho.

Contudo, para estimar variâncias e demais medidas derivadas (desvio-padrão, erro-padrão, valores de p etc.) é preciso usar rotinas especiais, que requerem, além do peso natural do desenho, a indicação das variáveis que contêm os códigos do estrato de seleção e da unidade primária de amostragem. No caso de São Paulo capital, cuja amostra conta com somente uma pessoa na faixa etária de 65-74 anos, o estrato de seleção tem apenas uma unidade primária de amostragem e é necessário agregá-lo com outro estrato de seleção para evitar problemas na estimação, que ocorrem em alguns sistemas estatísticos.

O uso do peso calibrado, entretanto, inclui outra fonte de aleatorização devida aos resíduos de calibração. Não se alteram os métodos para a obtenção de estimativas pontuais, mas, para as variâncias (e medidas derivadas), torna-se necessário considerar peso calibrado, códigos de estrato de seleção e unidade primária de amostragem e resíduos da regressão de calibração^{18,19,20,21}.

Alguns variáveis foram escolhidas para comparação da estatística descritiva amostral (sem intervalo de confiança – IC) com as estimativas obtidas, considerando os pesos naturais do desenho e pesos calibrados calculados. Nesse caso, foram usadas, respectivamente, as rotinas *surveymeans* e *surveyfreq* do SAS, que consideram a fonte de aleatorização do desenho de amostra complexo, e a biblioteca *survey*^{19,20}, versão 3.22.1, do R versão 2.11.1 (The R Foundation for Statistical Computing, Viena, Áustria; <http://www.r-project.org>), que considera as fontes de aleatorização devidas ao desenho complexo da amostra e à calibração dos pesos amostrais. Ressalte-se que estimativas obtidas com o peso calibrado são mais precisas do que as obtidas com o peso natural do desenho.

Resultados

A busca pelos relatórios e informações sobre a seleção da amostra da pesquisa foi realizada entre março de 2006 e dezembro de 2008 junto às coordenações das três esferas de governo e aos participantes da pesquisa que puderam ser identificados. Nessa busca, foram feitas mais de oitocentas ligações interurbanas, enviadas mais de mil mensagens eletrônicas e realizadas viagens a 14 UF.

As dificuldades mais comuns para estabelecer contatos (coordenações e participantes) foram decorrentes de dados errados, mensagens de correio eletrônico sem resposta e ligações telefônicas não atendidas. Nos casos de contatos estabelecidos, outras dificuldades surgiram: desconhecimento das coordenações municipais sobre a existência ou a localização dos relatórios e até sobre a pesquisa; retirada de dados e documentos do órgão pelos funcionários da gestão anterior; envio do material para o arquivo morto, Ministério da Saúde ou coordenadores regionais; não-fornecimento dos dados sem autorização formal do Ministério da Saúde. Nos casos cujo material foi arquivado em local conhecido, mas a coordenação (estadual/municipal) não dispunha de recursos humanos para sua busca, foi necessário viajar para realizar essa tarefa.

A busca, na COSAB, dos relatórios de pesquisa e da relação dos participantes com seus contatos não produziu os resultados esperados; por outro lado, comunicação com os coordenadores regionais permitiu recuperar relatórios das regiões Sudeste e Sul. Para as demais regiões, obteve-se informação de que os relatórios estavam perdidos. Os contatos com participantes das capitais, principalmente ex-coordenadores municipais, possibilitaram o resgate de relatórios ou da relação de nomes ou endereços dos estabelecimentos de ensino pesquisados.

Ao final, foi possível recuperar informação em 157 municípios do total de 250 selecionados, sendo 75 em 100 municípios (mais de 50 mil habitantes), incluindo as 27 capitais de UF, e 82 em 150 municípios (até 50 mil habitantes), cujas unidades de seleção são quadras, muitos deles contando com menos de vinte escolas (Tabela 1). Os relatórios obtidos indicaram situações não previstas nas instruções de seleção e coleta, conforme referido na seção de cálculo dos pesos amostrais para situações especiais.

O maior problema relacionado às variáveis de identificação do banco de dados do SBBrazil 2003 é a ausência daquelas que identifiquem todas as unidades de seleção. Esse entrave decorre do desenho incompleto da ficha de exame²²; faltam a identificação/numeração dos domicílios

Tabela 1

Tipo de informação recuperada segundo a macrorregião e o grupo de municípios.

Macrorregião e grupo de municípios	Municípios selecionados	Municípios com informação resgatada			
		Total *	Escolas **	Creches **	Setores **
Brasil					
Total de municípios	250	157	119	58	50
Até 50 mil habitantes	150	82	59	-	-
Mais de 50 mil habitantes (exceto capitais)	73	48	33	31	23
Capitais	27	27	27	27	27
Norte					
Total de municípios	50	27	22	9	8
Até 50 mil habitantes	30	16	13	-	-
Mais de 50 mil habitantes (exceto capitais)	13	4	2	2	1
Capitais	7	7	7	7	7
Nordeste					
Total de municípios	50	15	15	10	8
Até 50 mil habitantes	30	5	5	-	-
Mais de 50 mil habitantes (exceto capitais)	11	1	1	1	-
Capitais	9	9	9	9	8
Sudeste					
Total de municípios	50	50	37	16	8
Até 50 mil habitantes	30	30	19	-	-
Mais de 50 mil habitantes (exceto capitais)	16	16	14	12	4
Capitais	4	4	4	4	4
Sul					
Total de municípios	50	50	38	17	11
Até 50 mil habitantes	30	30	21	-	-
Mais de 50 mil habitantes (exceto capitais)	17	17	14	14	8
Capitais	3	3	3	3	3
Centro-oeste					
Total de municípios	50	15	7	6	14
Até 50 mil habitantes	30	1	1	-	-
Mais de 50 mil habitantes (exceto capitais)	16	10	2	2	10
Capitais	4	4	4	4	4

* Corresponde ao total de municípios para os quais foi obtida alguma informação sobre a seleção;

** Corresponde ao número de municípios para os quais a informação referida foi obtida.

e os códigos de creches/escolas. As identificações de unidades da amostra, constantes do banco, apresentam frequências importantes de dados faltantes ou inválidos. Do total de 108.921 registros do banco de dados, 27.619 (25,4%) pessoas que deveriam ser selecionadas em domicílios têm código de setor inválido ou ignorado, e 11.910 (10,9%) que deveriam ser selecionadas em creches/escolas têm código válido nas variáveis de distrito, subdistrito, setor e quadra/vila. De fato, os relatórios indicaram troca de cadastro de seleção, com escolares selecionados em seus domicílios, além de seleção nas faixas etárias de 18-36 meses e 15-19 anos fora dos domicílios.

Nas capitais, onde a troca de cadastro representa problema grave na definição da unidade primária de amostragem, houve apenas 63 registros com código de setor inválido para pessoas que deveriam ser pesquisadas em domicílio (de fato, o foram) e 1.651 registros de pessoas que deveriam ser selecionadas em escolas/creches com código válido de setor. Entretanto, as informações obtidas permitiram corrigir esses casos.

Outro problema observado relaciona-se à duplicação de valores na variável de identificação dos registros (IDENTIFI) e à duplicação de registros. Em princípio, a IDENTIFI deveria ser única por município, mas resultados mostram

que existem 2.102 registros nos quais essa variável não tem valor único (702 registros duplicados, 190 triplicados, cinco quadruplicados, seis quintuplicados e 13 com seis ocorrências). No exame da IDENTIFI por UF, município, sexo e idade, observaram-se 123 pares de registros idênticos e um par de registros com diferença nas variáveis de renda pessoal e familiar, em que um era a correção dos valores do outro (o corrigido foi escolhido).

Verificou-se, também, que o tamanho previsto da amostra não foi atingido para inúmeras combinações de município e grupo etário, além de não terem sido examinadas pessoas em outras: 18-36 meses (seis municípios); cinco anos (dois municípios); 12 anos (um município); 35-44 anos (um município); e 65-74 anos (três municípios).

Os resultados do procedimento de calibração foram sintetizados na Tabela 2, que compara os dados populacionais e estimativas obtidas com o peso natural do desenho e peso calibrado.

A Tabela 3 apresenta estatísticas das distribuições amostrais do peso natural do desenho e do peso calibrado, enquanto a Tabela 4 expõe as estatísticas descritivas da amostra, as estimativas para a prevalência de cárie, obtidas com o emprego dos dois pesos amostrais calculados, e a média dos índices de cárie e de seus componentes, para o total nacional de cada grupo etário. Ainda na Tabela 4, a fim de tornar os resultados comparáveis, buscou-se recuperar o algoritmo usado na publicação do Ministério da Saúde ¹, ignorando os códigos inválidos nas variáveis de resultado dos exames (DENTALXX, sendo XX = 11-18, 21-28, 31-38, 41-48): 37,9% de códigos numéricos (492.703 em 1.240.256 códigos) para dentes decíduos; 1,1% de códigos alfabéticos (24.488 em 2.245.216 códigos) para dentes permanentes.

A Tabela 5 apresenta as estatísticas descritivas da amostra publicadas e as estimativas obtidas com o emprego do peso natural do desenho e do peso calibrado, relativas às metas do milênio. Para as estimativas de prevalência incluiu-se o cálculo do total de pessoas que atendem às metas.

Discussão

Parte do trabalho de recuperação das informações sobre a seleção da amostra não seria necessária se o desenho da ficha de exame ²² tivesse forçado o registro de todas as unidades de seleção do desenho amostral, e não apenas os códigos de UF, município, setor e quadras ou vilas rurais. No entanto, esse trabalho permitiu concluir que replicar as instruções de seleção da amostra ⁸ traria viés ao cálculo dos pesos amostrais: os

resultados apresentados sobre as informações recuperadas indicam que a seleção da amostra nos municípios nem sempre foi feita de acordo com as regras estabelecidas. Mostrou, também, que o total de creches e escolas utilizado foi inferior ao existente no *Censo Escolar 2003*, forçando a introdução de ajustes no cálculo dos pesos amostrais para evitar subestimação importante nos grupos etários correspondentes.

Na crítica ao banco de dados da pesquisa, observou-se que parte dos problemas não tinha correção e parte foi alvo de medidas paliativas. Os casos de tamanho nulo da amostra efetiva não tinham solução. Para tamanho de amostra inferior ao previsto, fez-se correção nos pesos amostrais, apesar de conduzir a estimativas com precisão menor do que a prevista. Os problemas nas variáveis que identificam as unidades de pesquisa foram, quando possível e necessário, corrigidos para o cálculo dos pesos, mas não foram alterados na nova versão do banco de dados (com as informações estruturais do desenho amostral), visto ser impossível fazer uma correção abrangente. Os casos de duplicação da variável IDENTIFI, dentro dos municípios, foram contornados pela criação de uma variável de identificação sequencial dos registros (de 1 a 108.921), tornando possível gravar nestes as informações estruturais do desenho amostral, pressupondo-se que os casos referiam-se a pessoas diferentes e não a registros duplicados no processo de correção. Os 124 registros idênticos que estavam duplicados foram mantidos no arquivo para permitir a reprodução dos resultados anteriores, sendo criada uma variável para marcá-los.

Em face das dificuldades relatadas quanto à seleção da amostra e aos métodos usados no seu cálculo, os pesos naturais do desenho não produzem estimativas do total populacional dos estratos de seleção, por sexo e grupo etário, coerentes com as estimativas populacionais conhecidas por outros métodos (Tabela 2). Em princípio, quatro causas podem ser apontadas: as probabilidades de inclusão dos municípios foram calculadas usando os dados da estimativa populacional para 1º de julho de 1999 e a pesquisa foi realizada em 2002 e 2003; a seleção em creches/escolas não assegura que os matriculados sejam residentes no município e exclui da pesquisa os que não estejam matriculados; as probabilidades de inclusão das pessoas selecionadas em domicílios foram baseadas em quantitativos do *Censo Demográfico 2000*; comumente amostras domiciliares estimam a distribuição da população por sexo com vício ¹¹.

Por isso, o procedimento de calibração mostrou que, para os dados obtidos em domicílio, o peso natural do desenho tende a subestimar a

Tabela 2

Comparação entre dados da população e estimativas com pesos natural e calibrado, por sexo, segundo a faixa etária, a macrorregião e o grupo de municípios.

Faixa etária, macrorregião e grupo de municípios	Homens			Mulheres		
	Estimativa com peso natural	População	Estimativa com peso calibrado	Estimativa com peso natural	População	Estimativa com peso calibrado
18-36 meses						
Total	2.451.588,6	2.541.235	2.541.235,0	2.454.280,0	2.455.899	2.455.899,0
Norte	263.835,2	264.476	264.476,0	273.179,5	255.149	255.149,0
Capitais	43.534,7	43.290	43.290,0	37.457,6	42.055	42.055,0
Outros municípios	220.300,5	221.186	221.186,0	235.721,9	213.094	213.094,0
Nordeste	738.758,1	770.698	770.698,0	724.530,1	748.475	748.475,0
Capitais	73.796,6	66.096	66.096,0	47.670,5	63.833	63.833,0
Outros municípios	664.961,5	704.602	704.602,0	676.859,6	684.642	684.642,0
Sudeste	955.909,2	981.430	981.430,0	997.383,1	947.679	947.679,0
Capitais	182.350,0	234.587	234.587,0	266.752,3	225.767	225.767,0
Outros municípios	773.559,3	746.843	746.843,0	730.630,8	721.912	721.912,0
Sul	333.582,5	342.082	342.082,0	303.543,2	329.375	329.375,0
Capitais	40.674,5	40.162	40.162,0	34.695,0	38.943	38.943,0
Outros municípios	292.907,9	301.920	301.920,0	268.848,3	290.432	290.432,0
Centro-oeste	159.503,6	182.549	182.549,0	155.644,1	175.221	175.221,0
Capitais	54.767,3	63.620	63.620,0	65.695,2	61.138	61.138,0
Outros municípios	104.736,3	118.929	118.929,0	89.948,8	114.083	114.083,0
5 anos						
Total	1.067.039,5	1.826.306	1.826.306,0	1.063.905,3	1.763.738	1.763.738,0
Norte	119.998,3	182.888	182.888,0	122.725,3	175.456	175.456,0
Capitais	12.336,6	29.120	29.120,0	6.985,6	28.141	28.141,0
Outros municípios	107.661,7	153.768	153.768,0	115.739,8	147.315	147.315,0
Nordeste	395.163,0	563.713	563.713,0	398.208,7	544.763	544.763,0
Capitais	23.985,5	47.041	47.041,0	29.113,3	46.042	46.042,0
Outros municípios	371.177,5	516.672	516.672,0	369.095,4	498.721	498.721,0
Sudeste	341.686,9	695.617	695.617,0	332.476,8	673.770	673.770,0
Capitais	91.205,7	158.807	158.807,0	75.266,8	154.386	154.386,0
Outros municípios	250.481,2	536.810	536.810,0	257.210,0	519.384	519.384,0
Sul	141.324,8	254.057	254.057,0	145.249,2	243.498	243.498,0
Capitais	9.653,7	29.005	29.005,0	9.365,1	27.828	27.828,0
Outros municípios	131.671,1	225.052	225.052,0	135.884,1	215.670	215.670,0
Centro-oeste	68.866,5	130.031	130.031,0	65.245,4	126.251	126.251,0
Capitais	14.395,9	43.264	43.264,0	15.376,5	43.073	43.073,0
Outros municípios	54.470,7	86.767	86.767,0	49.868,9	83.178	83.178,0
12 anos						
Total	1.835.192,8	1.858.607	1.858.607,0	1.991.392,4	1.812.361	1.812.361,0
Norte	151.520,1	170.944	170.944,0	178.916,8	165.447	165.447,0
Capitais	40.337,3	25.352	25.352,0	35.316,8	25.299	25.299,0
Outros municípios	111.182,9	145.592	145.592,0	143.600,0	140.148	140.148,0
Nordeste	596.941,7	592.342	592.342,0	717.230,9	581.358	581.358,0
Capitais	46.441,4	50.590	50.590,0	48.110,0	51.659	51.659,0
Outros municípios	550.500,2	541.752	541.752,0	669.120,9	529.699	529.699,0
Sudeste	667.318,2	715.194	715.194,0	679.839,2	698.794	698.794,0
Capitais	184.655,1	160.607	160.607,0	153.584,0	159.700	159.700,0
Outros municípios	482.663,1	554.587	554.587,0	526.255,3	539.094	539.094,0

(continua)

Tabela 2 (continuação)

Faixa etária, macrorregião e grupo de municípios	Homens			Mulheres		
	Estimativa com peso natural	População	Estimativa com peso calibrado	Estimativa com peso natural	População	Estimativa com peso calibrado
Sul	270.561,5	252.478	252.478,0	266.285,2	243.378	243.378,0
Capitais	47.871,8	29.246	29.246,0	28.947,9	28.286	28.286,0
Outros municípios	222.689,7	223.232	223.232,0	237.337,3	215.092	215.092,0
Centro-oeste	148.851,3	127.649	127.649,0	149.120,2	123.384	123.384,0
Capitais	58.669,0	43.688	43.688,0	50.463,3	42.839	42.839,0
Outros municípios	90.182,2	83.961	83.961,0	98.657,0	80.545	80.545,0
15-19 anos						
Total	7.598.780,3	9.394.250	9.394.250,0	10.369.253,7	9.295.823	9.295.823,0
Norte	612.791,5	819.478	819.478,0	1.003.551,2	809.640	809.640,0
Capitais	70.045,3	133.926	133.926,0	192.455,5	142.371	142.371,0
Outros municípios	542.746,2	685.552	685.552,0	811.095,7	667.269	667.269,0
Nordeste	2.016.310,7	2.898.933	2.898.933,0	3.254.372,8	2.859.530	2.859.530,0
Capitais	189.280,2	255.861	255.861,0	318.264,4	284.447	284.447,0
Outros municípios	1.827.030,5	2.643.072	2.643.072,0	2.936.108,4	2.575.083	2.575.083,0
Sudeste	3.292.601,2	3.734.627	3.734.627,0	4.101.299,1	3.716.941	3.716.941,0
Capitais	850.603,1	881.130	881.130,0	890.415,4	910.464	910.464,0
Outros municípios	2.441.998,1	2.853.497	2.853.497,0	3.210.883,7	2.806.477	2.806.477,0
Sul	1.125.953,9	1.287.966	1.287.966,0	1.305.294,2	1.253.855	1.253.855,0
Capitais	162.523,9	163.195	163.195,0	143.303,1	164.220	164.220,0
Outros municípios	963.430,0	1.124.771	1.124.771,0	1.161.991,1	1.089.635	1.089.635,0
Centro-oeste	551.123,0	653.246	653.246,0	704.736,4	655.857	655.857,0
Capitais	241.551,9	239.500	239.500,0	239.604,7	256.475	256.475,0
Outros municípios	309.571,0	413.746	413.746,0	465.131,6	399.382	399.382,0
35-44 anos						
Total	7.287.635,7	11.536.482	11.536.482,0	15.161.698,6	12.229.912	12.229.912,0
Norte	436.769,7	773.854	773.854,0	1.022.754,8	744.473	744.473,0
Capitais	87.049,5	145.697	145.697,0	193.555,1	150.038	150.038,0
Outros municípios	349.720,2	628.157	628.157,0	829.199,7	594.435	594.435,0
Nordeste	1.622.151,0	2.732.428	2.732.428,0	3.476.578,4	2.973.803	2.973.803,0
Capitais	172.026,2	284.268	284.268,0	415.860,3	340.049	340.049,0
Outros municípios	1.450.124,8	2.448.160	2.448.160,0	3.060.718,1	2.633.754	2.633.754,0
Sudeste	3.521.681,9	5.325.754	5.325.754,0	7.023.394,4	5.691.185	5.691.185,0
Capitais	1.124.599,3	1.355.282	1.355.282,0	1.698.390,7	1.548.171	1.548.171,0
Outros municípios	2.397.082,7	3.970.472	3.970.472,0	5.325.003,7	4.143.014	4.143.014,0
Sul	1.213.025,5	1.877.146	1.877.146,0	2.519.509,7	1.963.004	1.963.004,0
Capitais	176.954,2	245.269	245.269,0	320.329,4	284.131	284.131,0
Outros municípios	1.036.071,3	1.631.877	1.631.877,0	2.199.180,2	1.678.873	1.678.873,0
Centro-oeste	494.007,6	827.300	827.300,0	1.119.461,4	857.447	857.447,0
Capitais	182.845,9	300.173	300.173,0	442.098,9	345.712	345.712,0
Outros municípios	311.161,7	527.127	527.127,0	677.362,4	511.735	511.735,0
65-74 anos						
Total *	2.127.484,5	2.979.835	2.797.556,0	4.026.568,0	3.589.465	3.589.465,0
Norte	111.241,6	165.580	165.580,0	184.281,7	158.233	158.233,0
Capitais	12.211,8	21.307	21.307,0	32.626,1	25.907	25.907,0
Outros municípios	99.029,7	144.273	144.273,0	151.655,5	132.326	132.326,0
Nordeste	537.519,4	784.058	784.058,0	1.076.124,9	937.245	937.245,0
Capitais	33.075,0	55.161	55.161,0	96.421,5	82.101	82.101,0
Outros municípios	504.444,4	728.897	728.897,0	979.703,4	855.144	855.144,0

(continua)

Tabela 2 (continuação)

Faixa etária, macrorregião e grupo de municípios	Homens			Mulheres		
	Estimativa com peso natural	População	Estimativa com peso calibrado	Estimativa com peso natural	População	Estimativa com peso calibrado
Sudeste *	940.959,6	1.377.582	1.195.303,0	1.974.314,5	1.735.511	1.735.511,0
Capitais *	139.200,7	361.956	179.677,0	709.034,6	528.627	528.627,0
Outros municípios	801.758,9	1.015.626	1.015.626,0	1.265.279,9	1.206.884	1.206.884,0
Sul	412.971,1	478.167	478.167,0	576.693,9	582.700	582.700,0
Capitais	67.432,0	59.678	59.678,0	73.413,8	90.316	90.316,0
Outros municípios	345.539,1	418.489	418.489,0	503.280,1	492.384	492.384,0
Centro-oeste	124.792,8	174.448	174.448,0	215.153,1	175.776	175.776,0
Capitais	44.052,2	52.255	52.255,0	69.313,3	65.281	65.281,0
Outros municípios	80.740,6	122.193	122.193,0	145.839,9	110.495	110.495,0

* Como na amostra do Município de São Paulo não houve entrevista de homem de 65-74 anos, há uma perda de 182.279 homens dessa faixa etária nas linhas de Capitais, Sudeste e Total do grupo etário, todas com valores em negrito.

Tabela 3

Parâmetros da distribuição amostral do peso natural do desenho e do peso calibrado.

Parâmetros	Peso natural do desenho	Peso calibrado
Número de observações no arquivo	108.921	108.921
Número de observações com valor válido *	108.797	108.797
Mínimo	1,7	4,8
Primeiro quartil (Q1)	42,3	52,5
Mediana	113,7	128,5
Terceiro quartil (Q3)	396,7	406,8
Máximo	431.460,5	259.745,0
Amplitude (máximo - mínimo)	431.458,8	259.740,2
Amplitude interquartilica (Q3-Q1)	354,3	354,3
Moda	87,9	33,4
Média	527,9	561,6
Desvio-padrão	2.366,7	2.228,6
Coefficiente de variação (%)	448,3	396,8

* Os 124 registros com valor duplicado ficaram com os pesos amostrais em branco.

população masculina e superestimar a feminina (fato comum em pesquisas domiciliares, pois as mulheres são mais facilmente encontradas em casa). O peso natural do desenho subestima a população de 18-36 meses e de cinco anos para ambos os sexos. Nas demais faixas etárias, o peso natural do desenho subestima a população masculina e superestima a feminina. Apesar das diferenças por sexo, o peso natural do desenho subestima a população total em todas as faixas etárias, exceto para 12 anos. Para a população total, os efeitos referidos não conduzem a uma compensação, e o peso natural do desenho ten-

de a subestimar a população total dos grupos considerados: a soma dos dados da Tabela 2 indica que a população total em 1^o de julho de 2003, para todos os grupos etários, corresponde a 61,3 milhões de habitantes, enquanto o peso natural do desenho fornece uma estimativa de 57,4 milhões.

Para os dados obtidos em creches (cinco anos, municípios com mais de 50 mil habitantes), o peso natural do desenho subestima o total, uma vez que muitas crianças dessa idade não estavam matriculadas nessas instituições (o total fornecido pelo *Censo Escolar 2003* é menor

Tabela 4

Estatísticas descritivas da amostra publicadas, estimativas baseadas no peso natural do desenho e estimativas obtidas com o peso calibrado para o total nacional de algumas variáveis, segundo o grupo etário.

Grupo etário e variáveis publicadas	Estatísticas descritivas da amostra publicadas *		Estimativas com peso natural do desenho			Estimativas com peso calibrado		
	Valor	%	Valor **	%	IC95%	Valor **	%	IC95%
18-36 meses								
Total	12.117	100,00	4.905.869	100,00		4.997.134	100,00	
ceo = 0	8.864	73,15	3.863.424	78,75	75,64; 81,87	3.940.863	78,86	75,99; 81,73
ceo ≥ 1	3.253	26,85	1.042.445	21,25	18,13; 24,36	1.056.271	21,14	18,27; 24,01
Média componentes ceo-d								
Hígidos	17,14		17,38		17,15; 17,61	17,36		17,15; 17,58
Cariados	1,03		0,82		0,65; 1,00	0,82		0,67; 0,97
Obturados e caritados	0,01		0,01		0,00; 0,01	0,01		0,00; 0,01
Obturados	0,03		0,03		0,01; 0,06	0,03		0,01; 0,06
Perdidos	0,01		0,01		0,00; 0,01	0,01		0,00; 0,01
ceo-d	1,07		0,87		0,68; 1,06	0,86		0,69; 1,03
5 anos								
Total	26.641	100,00	2.130.945	100,00		3.590.044	100,00	
ceo = 0	10.822	40,62	798.724	37,48	34,80; 40,18	1.453.934	40,50	38,50; 42,50
ceo ≥ 1	15.818	59,37	1.331.777	62,50	59,82; 65,20	2.135.643	59,49	57,48; 61,49
ceo com valor faltante	1	0,00	444	0,02		467	0,01	
Média componentes ceo-d								
Hígidos	16,05		15,54		15,17; 15,91	15,81		15,58; 16,04
Cariados	2,30		2,70		2,38; 3,02	2,38		2,16; 2,59
Obturados e caritados	0,06		0,05		0,04; 0,06	0,06		0,05; 0,07
Obturados	0,36		0,31		0,26; 0,36	0,34		0,30; 0,39
Perdidos	0,08		0,09		0,07; 0,11	0,08		0,06; 0,10
ceo-d	2,80		3,14		2,82; 3,47	2,86		2,64; 3,07
12 anos								
Total	34.550	100,00	3.826.585	100,00		3.670.968	100,00	
CPO = 0	10.739	31,08	1.316.344	34,40	32,02; 36,78	1.271.818	34,65	32,47; 36,82
CPO ≥ 1	23.811	68,92	2.510.242	65,60	63,22; 67,98	2.399.150	65,35	63,18; 67,53
Média componentes CPO-D								
Hígidos	22,70		22,98		22,76; 23,21	22,98		22,76; 23,20
Cariados	1,62		1,48		1,29; 1,66	1,46		1,31; 1,61
Obturados e caritados	0,07		0,06		0,05; 0,07	0,06		0,05; 0,07
Obturados	0,91		0,78		0,70; 0,86	0,79		0,71; 0,87
Perdidos	0,18		0,16		0,14; 0,19	0,16		0,14; 0,18
CPO-D	2,78		2,49		2,29; 2,68	2,47		2,29; 2,66
15-19 anos								
Total	16.833	100,00	17.968.034	100,00		18.690.073	100,00	
CPO = 0	1.861	11,06	2.555.744	14,22	12,37; 16,08	2.687.668	14,38	12,72; 16,04
CPO ≥ 1	14.972	88,94	15.412.290	85,78	83,92; 87,63	16.002.405	85,62	83,96; 87,28

(continua)

Tabela 4 (continuação)

Grupo etário e variáveis publicadas	Estatísticas descritivas da amostra publicadas *		Estimativas com peso natural do desenho			Estimativas com peso calibrado		
	Valor	%	Valor **	%	IC95%	Valor **	%	IC95%
Média componentes								
CPO-D								
Hígidos	22,24		22,91		22,59; 23,24	22,96		22,66; 23,26
Cariados	2,60		2,22		2,05; 2,38	2,24		2,07; 2,41
Obturados e cariadados	0,19		0,19		0,16; 0,22	0,19		0,16; 0,22
Obturados	2,49		2,42		2,19; 2,64	2,37		2,18; 2,57
Perdidos	0,89		0,72		0,65; 0,78	0,71		0,64; 0,77
CPO-D	6,17		5,54		5,24; 5,83	5,50		5,24; 5,77
35-44 anos								
Total	13.431	100,00	22.449.334	100,00		23.766.394	100,00	
CPO = 0	70	0,52	153.696	0,68	0,38; 0,99	170.306	0,72	0,41; 1,02
CPO ≥ 1	13.361	99,48	22.295.638	99,32	99,01; 99,62	23.596.088	99,28	98,98; 99,59
Média componentes								
CPO-D:								
Hígidos	10,85		11,36		10,88; 11,85	11,76		11,37; 12,15
Cariados	2,33		2,07		1,91; 2,23	2,19		2,06; 2,32
Obturados e cariadados	0,35		0,46		0,38; 0,53	0,45		0,38; 0,53
Obturados	4,22		5,22		4,82; 5,61	5,12		4,74; 5,50
Perdidos	13,24		11,68		11,12; 12,23	11,28		10,86; 11,70
CPO-D	20,13		19,40		18,90; 19,91	19,03		18,62; 19,44
65-74 anos								
Total	5.349	100,00	6.154.052	100,00		6.387.021	100,00	
CPO = 0	28	0,52	24.592	0,40	0,12; 0,68	25.981	0,41	0,12; 0,70
CPO ≥ 1	5.321	99,48	6.129.460	99,60	99,32; 99,88	6.361.040	99,59	99,30; 99,88
Média componentes								
CPO-D								
Hígidos	3,41		3,33		2,56; 4,09	3,63		3,04; 4,22
Cariados	1,17		0,93		0,74; 1,12	1,04		0,92; 1,16
Obturados e cariadados	0,06		0,06		0,04; 0,09	0,07		0,04; 0,09
Obturados	0,73		0,90		0,62; 1,17	0,91		0,70; 1,12
Perdidos	25,83		26,03		24,81; 27,24	25,58		24,83; 26,34
CPO-D	27,79		27,91		27,05; 28,78	27,59		27,00; 28,19

* As estatísticas descritivas da amostra correspondem às publicadas pelo Ministério da Saúde¹ (páginas 21, 24 e 26). Não foram apresentados intervalos de confiança para as estatísticas descritivas da amostra porque sua construção pressuporia que a amostra é autoponderada, hipótese que os dados da Tabela 3 indicam não ser realista, conforme Tabela 1 do artigo de Queiroz et al.⁴

** Na coluna valor, são apresentadas, para as linhas com a prevalência de cárie, a estimativa de total da população, ou, para as linhas com os componentes de ceo-d e CPO-D, as estimativas de médias.

Nota: O termo obturado não foi substituído por restaurado (preferido dos autores), para manter a terminologia usada na publicação do Ministério da Saúde¹. Os valores em negrito indicam estimativas cujo intervalo de confiança não contém o valor da estatística amostral publicada.

do que a estimativa da população dessa idade). Ademais, o número de creches usadas na seleção da amostra (casos cujas informações foram recuperadas) era menor do que no *Censo Escolar 2003*.

Na seleção em escolas (12 anos), os dados utilizados (não apresentados) mostraram que 94 municípios têm mais alunos de 12 anos matri-

culados do que residentes de 12 anos, ao passo que os 155 municípios restantes (um não teve amostra nesta idade) têm população superior ao número de alunos de 12 anos matriculados em escola. Os dados indicam que o peso natural do desenho só não superestima a população total de 12 anos nos outros municípios do Norte e Sudeste.

Tabela 5

Estatísticas descritivas da amostra publicadas, estimativas baseadas no peso natural do desenho e estimativas obtidas com o peso calibrado para as metas do milênio, segundo o grupo etário e a macrorregião.

Grupo etário e macrorregião	Estatística descritiva publicada *	Estimativas com peso natural do desenho			Estimativas com peso calibrado		
		Valor**	%	IC95%	Valor	%	IC95%
5 anos							
Meta do milênio		50% livres de cárie (ceo-d = 0)					
Brasil	40,56	798.724	37,48	34,79; 40,17	1.453.934	40,50	38,50; 42,50
Norte	35,04	79.872	32,91	24,91; 40,90	129.374	36,10	30,77; 41,44
Nordeste	34,92	256.797	32,37	26,73; 38,01	374.858	33,82	29,35; 38,28
Sudeste	44,82	304.223	45,13	41,86; 48,39	632.488	46,19	43,06; 49,31
Sul	43,26	106.848	37,28	31,92; 42,65	211.068	42,42	38,59; 46,26
Centro-oeste	41,73	50.985	38,02	30,57; 45,46	106.146	41,42	36,55; 46,29
12 anos							
Meta do milênio		CPO-D < 3,0					
Brasil	2,79	2,49		2,30; 2,68	2,47		2,29; 2,66
Norte	3,13	3,08		2,76; 3,40	3,07		2,81; 3,34
Nordeste	3,19	2,84		2,35; 3,32	2,84		2,35; 3,34
Sudeste	2,30	1,97		1,76; 2,18	1,99		1,78; 2,19
Sul	2,31	2,22		1,92; 2,51	2,23		1,95; 2,50
Centro-oeste	3,16	3,13		2,71; 3,56	3,18		2,79; 3,58
15-19 anos							
Meta do milênio		80% com todos os dentes na boca (P = 0)					
Brasil	61,09	11.693.206	65,08	62,08; 68,07	12.236.682	65,47	63,22; 67,73
Norte	49,26	791.852	48,99	44,50; 53,48	809.637	49,70	45,18; 54,22
Nordeste	53,05	2.919.554	55,39	51,40; 59,38	3.232.241	56,13	52,14; 60,12
Sudeste	71,22	5.363.713	72,54	65,66; 79,42	5.450.151	73,14	68,63; 77,66
Sul	70,96	1.786.194	73,47	70,39; 76,55	1.871.946	73,65	71,27; 76,02
Centro-oeste	65,68	831.893	66,24	61,85; 70,64	872.706	66,66	62,93; 70,40
35-44 anos							
Meta do milênio		75% com 20 ou mais dentes na boca					
Brasil	53,95	13.590.774	60,54	56,29; 64,79	14.806.109	62,30	59,39; 65,21
Norte	46,34	644.699	44,17	37,18; 51,16	722.621	47,59	42,39; 52,80
Nordeste	49,96	2.708.711	53,13	48,87; 57,38	3.176.298	55,66	51,75; 59,58
Sudeste	62,35	7.153.712	67,84	59,83; 75,85	7.598.992	68,98	63,25; 74,70
Sul	55,13	2.133.161	57,15	52,70; 61,6	2.272.454	59,18	55,59; 62,77
Centro-oeste	58,36	950.492	58,91	53,70; 64,11	1.035.745	61,48	56,76; 66,20
65-74 anos							
Meta do milênio		50% com 20 ou mais dentes na boca					
Brasil	10,23	650.239	10,57	7,66; 13,48	730.388	11,44	8,82; 14,05
Norte	8,58	18.898	6,39	3,07; 9,72	22.014	6,80	3,77; 9,83
Nordeste	11,07	179.705	11,14	8,41; 13,8	212.107	12,32	9,34; 15,30
Sudeste	9,32	311.169	10,67	4,70; 16,65	337.362	11,51	6,23; 16,79
Sul	10,42	106.625	10,77	7,66; 13,89	115.645	10,90	8,07; 13,73
Centro-oeste	11,22	33.843	9,96	5,67; 14,24	43.261	12,35	7,85; 16,86

* As estatísticas descritivas da amostra correspondem às publicadas pelo Ministério da Saúde¹ (p. 43). Não foram apresentados intervalos de 95% de confiança (IC95%) para as estatísticas descritivas da amostra porque sua construção pressuporia que a amostra é autoponderada, o que é uma hipótese que os dados da Tabela 3 mostram não ser realista, como já havia sido indicado pela Tabela 1 do artigo de Queiroz et al.⁴;

** Na coluna valor é apresentada a estimativa de total da população, exceto para o grupo etário de 12 anos onde é apresentada a média de CPO-D.

Nota: Os valores em negrito indicam estimativas cujo intervalo de confiança não contém o valor da estatística amostral publicada.

Esses fatos justificam o uso da calibração dos pesos amostrais, tendo em vista que as estimativas obtidas com o peso calibrado passaram a ser coerentes com a população para todos os segmentos (Tabela 2), exceto para homens do grupo de 65-74 anos. A inexistência de entrevistado no Município de São Paulo conduziu a uma subestimação de 182.279 pessoas da população, as quais ficaram sem representação na amostra.

A distribuição amostral do peso natural do desenho tem amplitude e dispersão muito grandes (Tabela 3). O maior peso é 253.800,3 vezes o menor peso, indicando que a amostra do SBBrazil 2003 está longe de ser autoponderada⁴. O coeficiente de variação do peso natural do desenho é muito grande (448,3%) e implica variabilidade nas estimativas derivadas dessa amostra, o que poderia ter sido evitado caso fosse adotado outro desenho de amostra. Kish²³ demonstra que os pesos reduzem o vício, mas aumentam a variância dos estimadores ponderados, visto que a razão entre a variância de um estimador ponderado de média e o correspondente estimador não ponderado é igual a 1 mais o quadrado do coeficiente de variação dos pesos amostrais.

Na realidade, a variabilidade dos pesos amostrais decorre do desenho da amostra e dos problemas observados no cálculo do tamanho da amostra. Este cálculo foi baseado em uma prevalência de cárie para os 12 anos e supôs um efeito de desenho de 2, quando este foi de 13,83 e 11,01, estimados, respectivamente, pelos pesos naturais do desenho e pesos calibrados.

O procedimento de calibração, apesar de reduzir a amplitude dos pesos em 39,8% e a relação entre maior e menor peso (o maior peso calibrado é 54.112,5 vezes o menor), teve efeito pequeno sobre o coeficiente de variação dos pesos: redução de 448,3% para 396,8%. Isso significa que, não obstante suas vantagens, o peso calibrado ainda implicará variabilidade nas estimativas maior do que o necessário, em virtude do desenho amostral empregado: a calibração não tem como evitar o desequilíbrio de tamanho de amostra entre os municípios grandes e os pequenos, já que o desenho foi concebido para ter aproximadamente o mesmo tamanho de amostra por município⁵.

Na Tabela 4, verifica-se que os intervalos de 95% de confiança (IC95%) para as estimativas da prevalência de cárie baseadas no peso natural do desenho não incluem o valor da estatística descritiva amostral, exceto a partir de 35 anos. A partir dessa idade, observa-se pequena variabilidade nessa variável em decorrência do efeito cumulativo do maior tempo de exposição à doença⁴. Em relação às médias dos índices (ceo-d/CPO-D), os IC95% das estimativas não incluem a

estatística descritiva amostral, com exceção para o grupo etário de 65-74 anos, que tem o menor tamanho amostral e, portanto, a amostra menos precisa. No que se refere a seus componentes, 14 dos 30 IC95% das estimativas não incluem a correspondente estatística descritiva da amostra.

Para as estimativas com o peso calibrado, a situação mantém-se semelhante, sendo observada a principal alteração na idade de cinco anos, provavelmente em decorrência da forte subestimação do peso natural do desenho nesse grupo (Tabela 4).

Em relação às metas do milênio (Tabela 5), as estimativas com peso calibrado indicam que: (1) não há alteração importante na meta para crianças com cinco anos de idade (mas a estimativa com o peso natural do desenho subestima, pelos motivos já indicados, a proporção de pessoas sem cáries); (2) a estatística descritiva da amostra superestima significativamente a média de cárie para a faixa etária de 12 anos; (3) a proporção de pessoas de 15-19 anos com todos os dentes é significativamente maior que a indicada pela estatística descritiva da amostra; (4) a proporção de pessoas de 35-44 anos com vinte ou mais dentes na boca é significativamente subestimada pela estatística descritiva da amostra; (5) não há alteração significativa na meta para o grupo etário de 65-74 anos, apesar do aumento de 1,2 ponto percentual na porcentagem de pessoas com vinte dentes ou mais na boca.

Em termos gerais, o levantamento das informações do SBBrazil 2003 mostrou problemas no registro e arquivamento de documentos da pesquisa. Muitos relatórios recuperados não seguiram o modelo estabelecido²², sugerindo que a atenção a esses documentos ficou aquém do indicado.

Apesar da precariedade dos dados em alguns relatórios e da não obtenção de relatórios de outros municípios, a estratégia desenvolvida viabilizou o resgate de informações referentes à seleção da amostra da pesquisa. Em adição, possibilitou a incorporação das informações estruturais do desenho amostral (estratos de seleção, unidades primárias de amostragem e pesos amostrais) ao banco de dados da pesquisa.

Os pesos calibrados, a despeito das aproximações de cálculo, permitem obter estimativas válidas e representativas da população de pesquisa e servem de base para a comparação histórica das condições de saúde bucal da população brasileira, a ser observada em novos estudos. Além disso, as informações estruturais do desenho da amostra, incorporadas ao arquivo de dados da pesquisa, permitem avaliar os erros de amostragem de forma até então não possível, contribuindo para melhorar a qualidade do ajuste de modelos

estatísticos e planejamento de futuras investigações. A calibração dos pesos amostrais mostrou-se importante para obter coerência com os dados populacionais, principalmente para a idade de cinco anos e correção da distribuição da população por sexo.

As estimativas apresentadas apontam que as estatísticas descritivas da amostra não refletem adequadamente a situação da saúde bucal no país à época da realização da pesquisa. Verifica-se que o Brasil encontrava-se em situação melhor em relação às metas do milênio do que foi divulgado pelo Ministério da Saúde¹.

São indiscutíveis os benefícios do SBBrazil para a odontologia brasileira e sua contribuição para o processo de construção do planejamento das ações odontológicas. Foi essa a percepção principal e motivadora deste trabalho, que se contrapôs às dificuldades enfrentadas ao longo de seu curso. Por isso, não se pode deixar de indicar os pontos positivos e negativos observados ao longo do presente estudo, a fim de contribuir para o aprimoramento do processo de pesquisa na área.

Logo, é recomendável que o treinamento continue focalizando a minimização dos erros de medida, mantendo o processo de calibração dos entrevistadores, sem deixar de concentrar-se nos cuidados necessários para evitar a introdução de vício nos estimadores em decorrência do não cumprimento das regras de seleção da amostra. Além disso, o desenho amostral deve ser concebido para evitar que os pesos amostrais variem de forma excessiva, de modo a controlar melhor a precisão das estimativas.

No fim do dia, o que importa é controlar o erro quadrático médio (função conjunta dos erros de medidas, dos vícios introduzidos no processo de amostragem e da variância amostral), sem perder de vista que, em pesquisas desse tipo, não seria esperado que os erros de medida dominassem o valor do erro quadrático médio.

Os autores reconhecem que a disseminação da versão do banco de dados do SBBrazil 2003 complementada pelas informações estruturais do desenho da amostra compete exclusivamente ao Ministério da Saúde; portanto, colocam ao dispor desta instituição os dados e documentação pertinente para seu uso, caso entenda que eles devam ser disseminados.

Resumo

Este é um artigo de amostragem probabilística cujo objetivo foi descrever os métodos usados para calcular e calibrar os pesos amostrais da Pesquisa sobre Condições de Saúde Bucal da População Brasileira (SBBrazil 2003) e identificar as demais variáveis estruturais do desenho da amostra. Apresenta uma síntese do trabalho de resgate das informações cadastrais usadas para seleção das unidades amostrais nos municípios amostrados. Descreve os principais problemas observados no banco de dados da pesquisa, que compuseram o conjunto de condicionantes para cálculo dos pesos naturais do desenho e para determinação das unidades primárias de amostragem e dos estratos de seleção. Por fim, compara algumas estimativas, obtidas por meio de pesos e demais variáveis estruturais da amostra, com as estatísticas amostrais descritivas publicadas, concluindo que as condições de saúde bucal da população brasileira eram melhores do que as divulgadas anteriormente.

Saúde Bucal; Amostragem; Condições de Saúde; Saúde Bucal

Contribuições

M. C. Portela, M. T. L. Vasconcellos e R. C. S. Queiroz conceberam o trabalho de recuperação das informações de seleção da amostra relativas aos participantes do SBBrazil 2003. R. C. S. Queiroz realizou o trabalho de contato com os participantes e sistematizou a informação recuperada. M. T. L. Vasconcellos e P. L. N. Silva estabeleceram as regras de cálculo dos pesos amostrais. M. T. L. Vasconcellos calculou os pesos amostrais e gerou o arquivo final. M. T. L. Vasconcellos e P. L. N. Silva escreveram o manuscrito, que foi revisado e aprovado por todos os autores.

Agradecimentos

Os autores agradecem a Angelo Giuseppe Roncalli da Costa Oliveira pelo fornecimento da listagem dos setores selecionados e pelas decisões sobre o tratamento dos registros duplicados no banco de dados do SBBrazil 2003. Este trabalho recebeu apoio da FAPERJ (processo E-26/171.336/2006), e três de seus autores são bolsistas do CNPq.

Referências

1. Departamento de Atenção Básica, Secretaria de Atenção à Saúde, Ministério da Saúde. Projeto SB Brasil 2003. Condições de saúde bucal da população brasileira, 2002-2003: resultados principais. Brasília: Ministério da Saúde; 2004.
2. Skinner CJ, Holt D, Smith TME. Analysis of complex surveys. Chichester: Wiley; 1989.
3. Narvai PC, Antunes JLF, Moysés SJ, Frazão P, Peres MC, Peres KG, et al. Validade científica de conhecimento epidemiológico gerado com base no estudo Saúde Bucal Brasil 2003. Cad Saúde Pública 2010; 26:647-57.
4. Queiroz RCS, Portela, MC, Vasconcellos MTL. Pesquisa sobre as Condições de Saúde Bucal da População Brasileira (SB Brasil 2003): seus dados não produzem estimativas populacionais, mas há possibilidade de correção. Cad Saúde Pública 2009; 25:47-58.
5. Criado IV, Cabria MSB. Procedimento de depuração de dados estadísticos. Vitória-Gasteiz: Instituto Vasco de Estadística; 1990.
6. Herzog TN, Scheuren FJ, Winkler WE. Data quality and record linkage. New York: Springer Publishing Company; 2007.
7. Instituto Brasileiro de Geografia e Estatística. Censo Demográfico 2000: agregado por setores censitários dos resultados do universo. 2ª Ed. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2003.
8. Área Técnica de Saúde Bucal, Departamento de Atenção Básica, Secretaria de Políticas de Saúde, Ministério da Saúde. Projeto SB2000. Condições de saúde bucal da população brasileira no ano 2000: manual do coordenador. Brasília: Ministério da Saúde; 2001.
9. Hájek J. Limiting distribution in sample random sampling from finite populations. Publications of the Mathematics Institute of the Hungarian Academy of Science 1960; 5:361-74.
10. Särndal CE, Swensson B, Wretman JH. Model assisted survey sampling. New York: Springer-Verlag; 1992.
11. Silva PLN. Calibration estimation: when and why, how much and how. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2004. (Textos para Discussão da Diretoria de Pesquisas, 15).
12. Vasconcellos MTL, Silva PLN, Szwarcwald CL. Sampling design for the World Health Survey in Brazil. Cad Saúde Pública 2005; 21 Suppl 1:S89-99.
13. Pessoa DGC, Silva PLN. Análise de dados amostrais complexos. São Paulo: Associação Brasileira de Estatística; 1998.
14. Kalton G. Compensating for missing survey data. Ann Arbor: University of Michigan, 1983.
15. Instituto Brasileiro de Geografia e Estatística. Metodologia das estimativas das populações residentes nos municípios brasileiros para 1º de julho de 2008. Uma abordagem demográfica para estimar o padrão histórico e os níveis de subnumeração de pessoas nos censos demográficos e contagens de população. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2008.
16. Silva PLN, Pessoa DGC, Lila MF. Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral. Ciênc Saúde Coletiva 2002; 7: 659-70.
17. Cochran WG. Sampling techniques. 3rd Ed. New York: John Wiley & Sons; 1977.
18. Hidiroglou MA, Rao JNK, Yung W. Variance computation for complex surveys using estimating equations. In: Proceedings of the Survey Methods Section. Annual Meeting of the Statistical Society of Canada. http://www.ssc.ca/survey/documents/SSC1999_M_Hidiroglou.pdf (acessado em 22/Out/2010).
19. Lumley T. Analysis of complex survey samples. J Stat Softw 2004; 9:1-19
20. Lumley TS. Complex surveys: a guide to analysis using R. New York: John Wiley & Sons; 2010.
21. Chambers RL, Skinner CJ. Analysis of survey data. Chichester: John Wiley & Sons; 2003.
22. Área Técnica de Saúde Bucal, Departamento de Atenção Básica, Secretaria de Políticas de Saúde, Ministério da Saúde. Projeto SB2000. Condições de saúde bucal da população brasileira no ano 2000: manual do coordenador. Brasília: Ministério da Saúde; 2001.
23. Kish L. Weighing for unequal Pi. J Off Stat 1992; 8:183-200.
24. Kish L. The hundred years' wars of survey sampling. In: Proceedings of the 1995 Rome Conference to Commemorate 100 Years of Sample Survey. Rome: Centro d'Informazione e Stampa Universitaria; 1996. p. 15-27.

Recebido em 16/Out/2009

Versão final reapresentada em 30/Jul/2011

Aprovado em 18/Out/2011