CAPÍTULO 16

MÉTODOS PARA A AVALIAÇÃO DA QUALIDADE DE DADOS

16.1 INTRODUÇÃO

Nas palavras de Jenkinson (2002), "No census of population could ever claim to be error-free". Como acontece com qualquer dado social, os dados demográficos, dependendo da fonte da qual provêm, estão sujeitos a diferentes tipos de erros. Os dados censitários, devido ao tamanho da operação censitária e a dificuldade do treinamento e da supervisão de um número de entrevistadores que pode chegar a centenas de milhares em alguns países, está particularmente sujeito a erros. Erros estão presentes em todo tipo de levantamento de informação, mas a demografia está particularmente atenta à sua detecção e correção. Por um lado, isso acontece porque a informação demográfica é muito básica e serve de ponto de partida para muitos outros levantamentos de dados. Se uma pesquisa de mercado superestima a demanda por um determinado produto em 5% ou se uma pesquisa de opinião pública exagera a popularidade do governador do Estado em 5%, as consequências geralmente não serão graves. Mas se o censo de população deixa de enumerar 5% da população de um município, isso tem consequências reais no orçamento no municipal e no planejamento dos serviços essenciais. A outra razão pela qual a demografia tende a preocupar-se mais com os erros e a sua correção é a natureza dos dados demográficos, que tipicamente contém certas regularidades e redundâncias que permitem avaliar com mais facilidade quais são as inconsistências do que acontece com outros tipos de dados. Por exemplo, se o censo conta 50.000 crianças de 0 anos e o registro civil do último ano só registrou 35.000 nascimentos, pode-se desconfiar de que pelo menos uma das fontes está errada.

No Capítulo 4 introduziram-se os termos *erro de cobertura* e *erro de conteúdo*, no contexto do censo demográfico, mas estes termos têm uma aplicabilidade mais ampla. No caso do exemplo do parágrafo anterior, a inconsistência entre os dados do censo e do registro civil pode indicar um erro de cobertura do registro civil, no sentido de sub-registrar o número de nascimentos, ou pode ser um erro de cobertura do censo, no sentido oposto de superenumerar a população de crianças de 0 anos. Erros de superenumeração são menos comuns do que erros de subenumeração, mas podem acontecer, talvez porque algumas crianças foram enumeradas mais do que uma vez ou porque algumas crianças enumeradas já tinham mais de 1 ano de idade. Neste último caso, o erro em realidade seria de conteúdo e não estritamente de cobertura, já que foi o conteúdo do dado recolhido (idade) que causou o problema.

Os erros mais comuns encontrados em dados demográficos, divididos em erros de cobertura e de conteúdo, são os seguintes:

Erros de cobertura

1. Subenumeração ou superenumeração do censo demográfico por motivos políticos, para privilegiar certos grupos étnicos ou entidades políticas. Em alguns países onde existem conflitos étnicos, os censos são uma operação politicamente muito controvertida que sempre evoca grandes suspeitas de que um ou outro grupo esteja tentando exagerar a sua própria importância numérica ou diminuir a importância de outros grupos.

2. Subenumeração ou superenumeração censitária de certas categorias de pessoas que são mais difíceis de contar como crianças recém-nascidas ou populações com alto grau de mobilidade. As crianças recém-nascidas muitas vezes são omitidas por esquecimento ou porque ainda não são consideradas plenamente como membros da família. As populações com alto grau de mobilidade, como migrantes sazonais ou moradores de rua, correm o risco de serem omitidas na contagem ou então podem ser enumeradas mais de uma vez. Também existe a possibilidade de subenumeração diferencial por sexo em alguns grupos etários como homens jovens.

3. Sub-registro de nascimentos e/ou óbitos: O problema neste caso é a omissão por parte da população em registrar tais eventos, seja porque o registro tem um custo, porque as entidades de registro são distantes ou porque não há nenhum benefício aparente em fazer o registro. Como se mostrou no Capítulo 5, o Brasil demorou muito tempo em conseguir um registro de nascimentos mais ou menos completo. Em outros países os problemas maiores existem em relação ou registro de óbitos.

Erros de conteúdo

1. Erros na declaração da idade. Aqui pode-se distinguir entre diversos tipos de erros. Um tipo é a chamada *preferência digital* (“digit preference” ou “age heaping”, em inglês) a tendência a arredondar as idades para números que terminam em determinados dígitos como “0” ou “5”. Outro erro é que as pessoas podem evitar ou – pelo contrário – ser atraídas por certas idades específicas como a idade que lhes dá o direito de dirigir ou uma faixa que já

seria considerada "velha" (a passagem dos 30 anos) ou uma idade limite que implica em responder perguntas adicionais do questionário. As pessoas mais velhas tendem muitas vezes a exagerar a sua idade. E finalmente existem sistemas tradicionais de contagem da idade que são diferentes dos sistemas ocidentais.

2. Erros de memória. Em perguntas que exigem que o respondente se lembre de fatos ocorridos no passado remoto, sempre existe o risco do esquecimento. Por exemplo, mulheres mais velhas que precisam relatar quanto filhos nascidos vivos tiveram ao longo das suas vidas tendem a esquecer filhos nascidos há mais tempo, principalmente se morreram quando jovens. Em outros casos, certos eventos podem ser subdeclarados porque o respondente não gosta de lembrá-los, por exemplo óbitos recentes de membros do domicílio (agregado familiar).

3. Erros de referência de tempo. Algumas perguntas exigem que o respondente se lembre quantos eventos ocorreram durante um determinado período como o último ano ou o mês passado. A experiência ensina que as pessoas muitas vezes não dimensionam estes períodos corretamente e acabam incluindo eventos que ocorreram fora do prazo ou excluindo eventos que ocorreram dentro dele. Este tipo de erro dificulta, por exemplo, a estimação da fecundidade recente das mulheres.

4. Erros de preenchimento. Os entrevistadores do censo e – em menor medida – de inquéritos às vezes deixam respostas em branco. Rigorosamente isso deveria acontecer só em casos onde a pergunta não se aplica ou o entrevistado não respondeu. Mas muitas vezes também acontece porque o entrevistador considerou a resposta óbvia ou porque o número solicitado foi 0. Por exemplo, um entrevistador pode considerar que uma menina de 16 anos que mora com os pais "obviamente" deve ser solteira e não ter filhos. Mas ao deixar os campos em branco, estes serão codificados como "sem informação" e não como "solteira" e "0 filhos". Para o problema específico de números de filhos não declarados existe uma técnica de correção, conhecida como a técnica de El Badry, que será discutida no Capítulo 23.

5. Mesmo que toda a informação solicitada seja fornecida corretamente pelos entrevistados, certas perguntas têm limitações estruturais que não permitem a captação de todas as tendências demográficas relevantes. Isso acontece particularmente com as perguntas sobre migrações. Como já se viu no Capítulo 11, por exemplo, a pergunta sobre a residência do respondente numa data fixa do passado fornece informação sobre o resultado acumulado dos diferentes movimentos que ocorreram durante o período, mas não consegue captar todas as etapas intermédias e especificamente não detecta eventuais movimentos circulares.

Esta lista identifica os erros mais comuns do ponto de vista da medição dos processos demográficos básicos. Mas existem vários outros relacionados com temas mais específicos como a atividade econômica. Por exemplo, há uma tendência reconhecida por parte de pessoas que não têm um emprego regular e não possuem a sua própria empresa a declarar que "não trabalham", mesmo quando realizam atividades diversas irregulares. No caso da medição da deficiência, uma pessoa pode não reconhecer a sua condição como uma deficiência, principalmente quando a condição é de longa data e tanto a pessoa como o seu ambiente convivem com ela com naturalidade. E

no que toca à migração internacional há uma tendência óbvia a não declarar o status de imigrante por parte de pessoas que estão no país em condição irregular.

Dada a variedade de erros que podem ocorrer e a variedade de soluções para corrigi-los, este capítulo foca numa categoria particularmente fundamental e que em muitos casos pode ser corrigida, que são os erros na enumeração por sexo (segundo ponto da lista de erros de cobertura) e na declaração da idade (ponto 1 da lista de erros de conteúdo). Algumas das outras categorias da mesma lista (particularmente 2), 3) e 4) dos erros de conteúdo) serão abordadas no Capítulo 23.

Como se viu no Capítulo 6, a informação básica para iniciar uma análise demográfica depende muito das variáveis sexo e idade. Da composição de ambas é possível fazer uma primeira e muito robusta análise, tanto da atual composição da população por sexo e idade como da passada e ainda, futura. Adicionalmente, a desagregação da informação por sexo e idade – e principalmente esta última – cumpre um papel central no cálculo de todos os indicadores demográficos. Mais especificamente, grande parte da racionalidade do instrumental analítico demográfico se apoia na composição etária da população para a medição e interpretação dos fenômenos demográficos e suas tendências passadas e futuras. Por estas razões é necessário, antes de proceder a qualquer análise, avaliar a confiabilidade que a informação sobre sexo e idade possa ter, considerando a maior quantidade possível de aspectos envolvidos. Não se deve descuidar, por exemplo, o tipo de formulação das perguntas, as instruções dadas ao entrevistador, o tratamento dos casos de não resposta ou a inconsistência com outras características do entrevistado.

Mesmo assim, os erros na declaração da idade continuam sendo um dos problemas mais frustrantes da demografia (Ewbank, 1981: 88), o que tem forçado o demógrafo a desenvolver métodos para avaliar a qualidade dos dados de idade desde os primórdios da sistematização do conhecimento demográfico. Este item se inspira fortemente no importante e exaustivo trabalho realizado pelo mencionado autor para a realidade dos anos 80. Em que pese a evolução dos mecanismos de recolha de dados e do seu processamento automático, muitos dos aspectos por ele levantados continuam válidos.

Este capítulo apresenta o perfil e padrões esperados da declaração da informação sobre sexo e idade, assim como uma noção introdutória sobre a qualidade da mesma; descreve em primeiro lugar, o que se espera da informação sobre sexo e o potencial significado dos desvios do que é esperado. Em segundo lugar, analisa as possibilidades de erro na declaração da idade e alguns procedimentos que permitem avaliar o grau de confiabilidade destas duas variáveis seja por separado, seja conjuntamente. A menção a procedimentos de ajuste ou correção desta informação é feita de forma a orientar etapas posteriores se houver necessidade de maior aprofundamento nas análises. Um instrumental mais complexo para estas etapas posteriores pode ser encontrado, por exemplo, em Moultrie et al. (2013: Cap. 1).

16.2 IRREGULARIDADES NA DISTRIBUIÇÃO POR SEXO

Salvo fenômenos especificamente diferenciais por sexo, há um relativo consenso histórico sobre a composição populacional por sexo relativamente constante; Louis Henry (1948) já constatou esta constância e uma revisão posterior da literatura por Chahnazarian (1988) a confirma. A razão desta estabilidade se deve, basicamente, à condição orgânica ou biológica que determina uma proporção por sexo relativamente constante ao nascer (ver Capítulo 6) e a uma mortalidade

por idade, no geral, maior para os homens. No passado recente encontrou-se, com alguma frequência, maior mortalidade feminina devido, basicamente, a causas de morte ligadas à gravidez e que hoje têm menor incidência. Semelhante diferenciação tem se apresentado, mais recentemente, em sociedades que segregam as mulheres (ver Capítulo 7; Guilmoto, 2009). Dada esta regularidade esperada, desvios evidentes das Razões de Sexo ao longo do tempo ou por idade podem ser considerados indícios de erros na informação.

16.3 IRREGULARIDADES NA DISTRIBUIÇÃO POR IDADE

Numa população fechada à migração, ausente de conflitos e/ou mudanças bruscas, é de se esperar que a cada ano, surja uma geração –ou coorte– cujo tamanho seja similar às mediatamente anteriores ou posteriores, respeitando as tendências de crescimento positivo ou negativo que tal população possa ter. Consequentemente, é de se esperar que ao longo do tempo, a distribuição por idades dessa população siga um comportamento suave, acorde a esse crescimento. No entanto, a declaração da idade, como qualquer variável captada numa pesquisa, está sujeita a erros que são passíveis de detectar e eventualmente ajustar. Segundo Coale (em Ewbank, 1981: xiv), isto é possível, devido, em parte, às peculiaridades da variável idade como são:

- Aumento linear com o tempo, o que oferece enormes possibilidades de modelagem;
- Conhecimento geralmente amplo sobre a idade, baseado, por exemplo em certificados e rituais e atitudes além do que, como norma, a idade de uma pessoa é mais conhecida por outros membros da família do que outras variáveis e, portanto, pode ser mais facilmente declarada mediante proxies;
- Alta correlação com características físicas do corpo;
- Menos sensibilidade social que outras variáveis (aborto, crime, contracepção, riqueza ou renda, atitudes) sendo, portanto, menos sensível às limitações dos processos de recolha do dado;
- É uma medida objetiva, o que não se aplica, por exemplo, a atitudes comportamentais.

Considerando o contexto do século XXI, a seguir, este item compreende: as fontes de erros de declaração da idade e as técnicas disponíveis para identificar e medi-los; algumas técnicas usadas para captar distorções na declaração da idade devidas a erros e/ou subenumeração; possibilidades do uso do referencial das populações teóricas e das tábuas de mortalidade (ou sobrevivência).

16.3.1 Como é recolhida a informação sobre idade

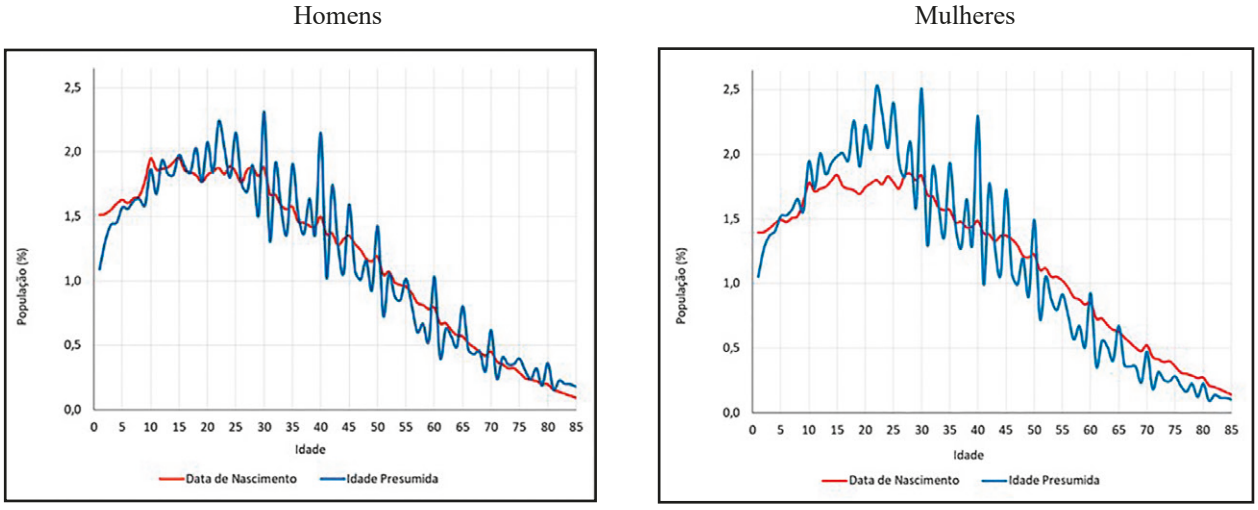
Como já se mencionou no Capítulo 4, a idade é captada basicamente de duas formas:

- Data de nascimento, com especificação de, pelo menos, mês e ano e eventualmente o dia do nascimento.
- Idade em anos completos (ou por completar como é o caso de vários países asiáticos), na ausência de resposta no caso anterior.

Teoricamente ainda há uma terceira possibilidade, a saber o uso do calendário histórico (ver Capítulo 8), mas esta técnica só pode ser usada em pesquisas antropológicas de pequeno porte e não serve para censos nacionais. Na maioria dos países o censo pergunta a idade das pessoas. O censo português pergunta a data de nascimento. O censo brasileiro pergunta tanto a idade como a data de nascimento.

A generalidade dos questionários censitários e da maioria dos inquéritos amostrais inclui ambas formas, sendo o entrevistador instruído a privilegiar a primeira alternativa e, como norma, pedir documento de identidade ou comprobatório da data do nascimento. Esta prática permite controlar ambas respostas, dá mais confiabilidade aos dados e elementos para, eventualmente, corrigi-los. Está comprovado que a variável idade, recolhida via a data de nascimentos costuma ser menos errática que aquela recolhida via a simples declaração da idade. O Gráfico 16.1, com dados do Censo brasileiro de 2010, evidencia esta afirmação e permite ver que as oscilações são mais acentuadas ainda, no caso das mulheres.

Gráfico 16.1: População por idades simples segundo idade estabelecida mediante a data de nascimento ou idade presumida - para homens e mulheres, Brasil - 2010



Fonte: IBGE - SIDRA (https://sidra.ibge.gov.br/acervo#/S/Q).

Ao contrário da variável sexo, a resposta sobre a idade está mais sujeita a erros originados tanto pelo desconhecimento da informação como pela deficiente apuração do dado e ainda motivados por fatores basicamente culturais. O manual de recomendações e princípios sobre formulação de censos (United Nations, 2017 a) detalha as possíveis fontes de erros e a forma de evitá-los no momento da colheita. Com tudo, uma etapa importante na fase de avaliação da declaração da idade é o conhecimento das instruções dadas ao entrevistador sobre a formulação da pergunta e dos procedimentos para resolver situações ambíguas, como por exemplo: se houve tendência a responder pela data do último aniversário, do seguinte aniversário ou o mais próximo aniversário; se há referência a mais de um sistema calendário (sistema lunar, ocidental etc.) e se foram dadas instruções para fazer as conversões; como se recolhe a informação para os menores de um ano ou para os mais idosos. Qual foi o tipo de informante no domicílio (agregado familiar) que mais frequentemente respondeu pelos outros membros. Nas situações de alta incidência de idade ignorada é necessário conhecer

que parâmetros ou referências foram usados para eventuais imputações. O predomínio da resposta 'idade ignorada' pode ter origem no pouco valor que culturalmente outorga-se à precisão para quantificar a idade, o que costuma se dar nas sociedades tradicionais. O culto à juventude e beleza feminina em sociedades ocidentais modernas explicariam a tendência das mulheres a declararem menos anos dos que realmente têm. A valoração da velhice em outros contextos, ou o deterioro físico da pessoa em ambientes hostis (de pobreza, inclemências climáticas, por exemplo) costumam justificar a adjudicação de idades superiores às reais entre as pessoas idosas.

É importante considerar, por último, que, corriqueiramente, a análise a ser feita é a idade e não a data de nascimento, pois é assim como, geralmente divulgam-se os dados.

Entre os padrões típicos de erro na informação sobre idade completa estariam:

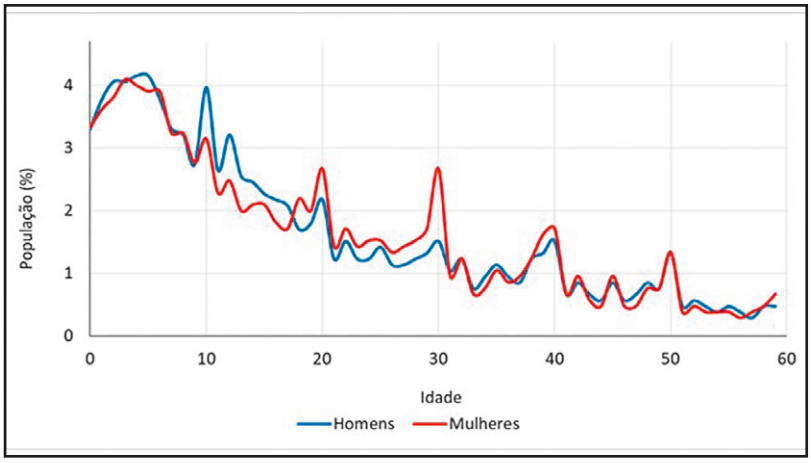
- Preferência para o arredondamento a dígitos 0, 5, 2 e 8 (usualmente nesta ordem);
- Omissão dos menores de um ano de idade – e frequentemente aplicado ao grupo de 1-4 anos);
- Tendência ao aumento da idade das pessoas mais idosas;
- Melhor qualidade da resposta entre a população masculina.

A inspeção da distribuição por idades simples numa população é o primeiro indicador da qualidade desta informação.

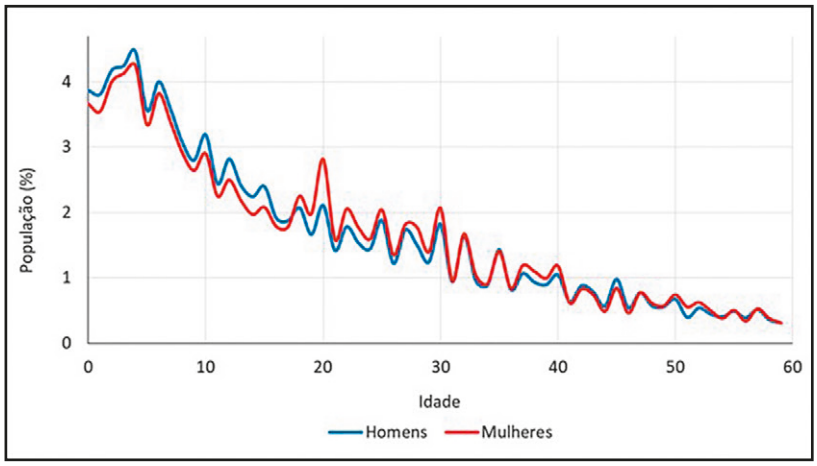
Os Gráficos 16.2.A e B, baseados em informações censitárias, mostram a declaração por idade e sexo para Moçambique em dois momentos (1980 e 2009) e o Gráfico 16.2.C, para o Brasil em 2010 (embaixo). Em maior medida para primeiro momento, os padrões de erro citados estão presentes. O menor percentual de crianças, em ambos sexos, tanto pode ser simples omissão das mesmas como uma real diminuição da natalidade.

Gráfico 16.2: Distribuição relativa da população por idades simples (homens e mulheres) Moçambique, 1980 e 2009 e Brasil, 2010

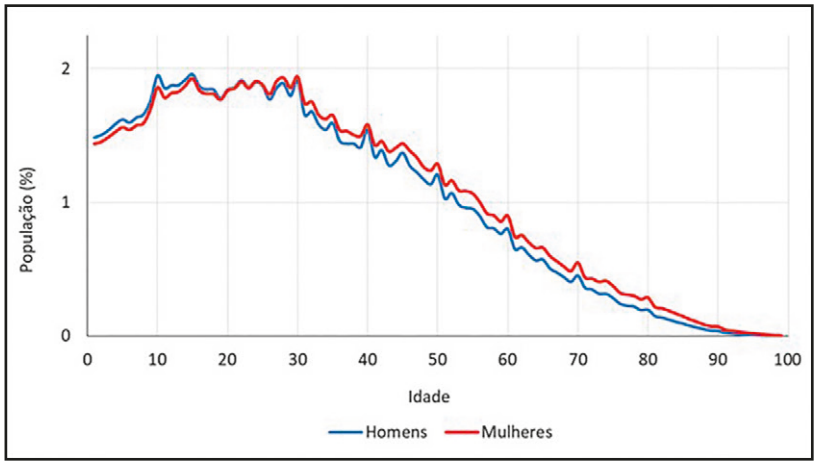
A. Moçambique, 1980



B. Moçambique, 2009



C. Brasil, 2010



Fontes: Gaspar (1989); 2009; UNSD Demographic Statistics, United Nations Statistics Division. http://data.un.org/Data.aspx?d=POP&f=tableCode%3A22; IBGE - SIDRA (https://sidra.ibge.gov.br/acervo#/S/Q).

Dada a experiência histórica dos censos, é provável que a omissão seja a causa mais importante nesta menor porcentagem de crianças menores de um, dois até quatro anos, principalmente para o contexto demográfico de Moçambique dos anos 80. As fortes oscilações ao longo das outras idades indicam o deslocamento da idade a valores vizinhos, assim por exemplo, pessoas que preferiram declarar a idade 25, seriam provavelmente aquelas com idades reais de 23, 24, 26 ou 27, que, no gráfico, mostram-se desfalcadas. A preferência pelas idades 12 e 22, nos dois momentos, são um bom exemplo.

Uma distribuição de idade mais regular depende, não apenas de bom planejamento, execução e recolha de dados. Depende também, como dito, do grau de importância que esta variável tem para cada indivíduo, o que por sua vez depende do tipo de sociedade em que ele está inserido. Se a idade é um dado ignorado, o investimento numa boa qualificação do entrevistador e em qualquer aspecto do censo ou pesquisa terá pouco efeito. A evolução socioeconômica experimentada por

Moçambique durante o período considerado explicaria, tanto o maior esforço por uma melhor recolha dos dados como a melhora na declaração da idade. Contextos mais desenvolvidos, como o caso do Brasil, em 2010, dentro dos PALOP, têm conseguido recolher esta informação de forma relativamente mais confiável. No caso do Gráfico 16.2.C, que mostra os dados do Censo brasileiro de 2010, ainda há certa preferência pelo arredondamento nas idades terminadas em 0 e que se nota mais na população feminina. De qualquer forma, as oscilações são menos acentuadas em relação às distribuições apresentadas em 16.2.A e 16.2.B.

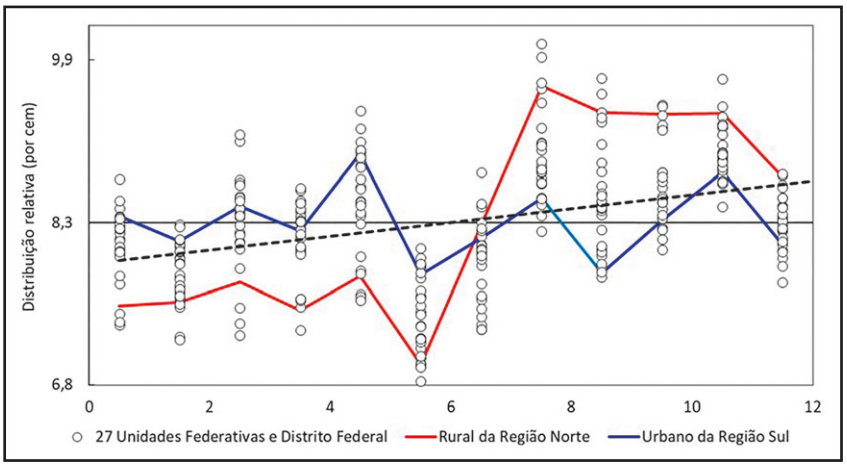
16.3.2 A declaração da idade de menores de um ano

No caso dos menores de um ano, em que se registra a idade em meses completos, a distribuição das idades – na ausência de fatores sazonais que afetem a natalidade ou a sobrevivência dos menores de um ano – a distribuição da idade deveria oscilar em torno de 1/12 (ou 8,3%) se declarada corretamente. O caso do Brasil serve para ilustrar essa distribuição. O Gráfico 16.3.A mostra os desvios com relação a uma distribuição uniforme de crianças de menores de um ano segundo idade em meses nas 28 divisões administrativas (as 27 Unidades Federativas e o Distrito Federal). Inclui, também a área urbana da Região Sul e a área rural da Região Norte, estas duas como aproximação de contextos mais e menos desenvolvidos respectivamente.

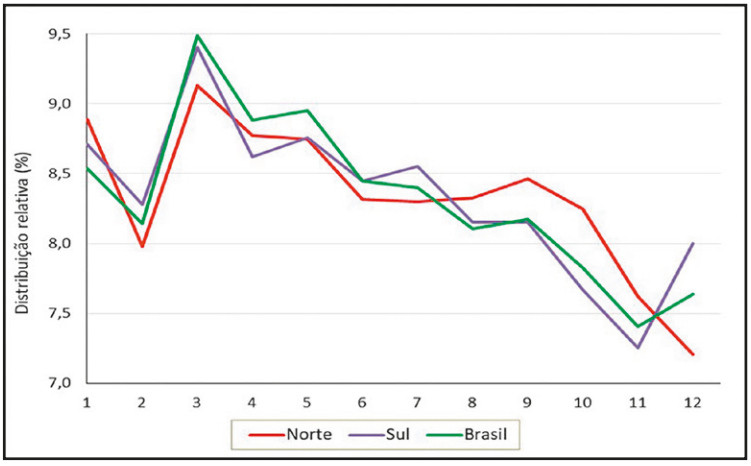
A distribuição correspondente às unidades administrativas indica, a julgar pela tendência linear que elas definem (linha pontilhada), certa concentração de infantes no segundo semestre de vida. No caso do contexto mais desenvolvido, embora se note uma preferência por idade em meses pares, o formato zigue-zague em torno do 8,3% denota desvios relativamente menores. Já no contexto menos desenvolvido (Rural da Região Norte), de forma oposta, há uma muito acentuada concentração de crianças no segundo semestre; neste caso, há de se procurar uma explicação, podendo levantar algumas hipóteses a pesquisar. Importante lembrar, antes, que a idade declarada se relaciona com a data de referência do censo, que no Brasil foi 31 de julho; assim: haveria fatores sazonais que provoquem nascimentos ocorrendo com relativamente mais frequência no primeiro semestre do ano, logo os mais velhos se localizariam no segundo semestre? Porque este fenômeno apresenta-se no contexto menos desenvolvido? Há algum elemento cultural que faça aumentar a idade declarada das crianças menores de um ano? Há algum padrão de erro associado ao nível de desenvolvimento ?

Gráfico 16.3: Declaração segundo meses de idade ou do nascido vivo – Brasil, 2010

A. Distribuição relativa da idade em meses dos menores de um ano – Urbano, região Sul e Rural, região Norte



B. Distribuição relativa do mês de ocorrência dos nascidos vivos (Brasil, e regiões Norte e Sul)
Média dos anos 2009 a 2011



Fonte: IBGE - SIDRA (https://sidra.ibge.gov.br/acervo#/S/Q e https://sidra.ibge.gov.br/tabela/2612).

No intuito de verificar se há algum padrão de erro associado ao nível de desenvolvimento, avaliar a confiabilidade desta informação, utilizando fontes externas é aconselhável. Isto pode ser feito utilizando, por exemplo, o Registro Civil que costumeiramente disponibiliza o mês de ocorrência do nascimento. O Gráfico 16.3.B, apresenta a distribuição relativa por meses deste evento para o período 2009 a 2011, como forma de evitar eventuais oscilações, mas representativa do ano 2010. Embora não estritamente comparável à desagregação apresentada no gráfico anterior, pois neste caso as estatísticas se referem aos totais regionais, observa-se que, igual que no caso da informação censitária, a distribuição não é uniforme. As estatísticas contínuas confirmam que, efetivamente, haveria uma relativa maior concentração de nascimentos no primeiro semestre do ano. A coincidência registrada outorga, desta forma, confiabilidade a ambas as fontes.

16.3.3 Índices para quantificar a irregularidade da distribuição

Existem muitas formas de avaliar a declaração da idade, assim como propostas de correção ou redistribuição. No caso da avaliação, há uma série de índices que existem desde os anos 50 para medir a preferência/rejeição de cada dígito ou de alguns em especial (United Nations, 1955). Sua aplicação possibilita a comparação da qualidade entre duas fontes e assim implementar estratégias de melhora na colheita do dado. Em breve, se trata dos seguintes procedimentos:

Índice de Whipple

O índice proposto por Whipple (Whipple, 1919; Siegel e Swanson, 2004; Nações Unidas, 1955; United Nations, 2017 a) calcula a porcentagem devida à soma das idades que terminam com 0 e 5 em relação a 1/5 da soma dos valores de idades de 23 a 62 anos. Na ausência de qualquer preferência por números que terminam em "0" ou "5", o índice é igual a 1 (ou 100, se for multiplicado por 100). No outro extremo, se todas as idades declaradas terminarem em "0" ou "5", o seu valor será 5 (ou 500).

$$W_{0\,e\,5} = 500 \cdot (P_{25} + P_{30} + P_{35} + P_{40} + P_{45} + P_{50} + P_{55} + P_{60}) \,/\, {}_{40}P_{23} \text{ (intervalo 23 – 62)} \qquad (16.1)$$

Os resultados geralmente são interpretados da seguinte forma:

Menos de 105	Dados muito exatos
105-110	Dados relativamente exatos
110-125	Dados aproximados
125-175	Dados grosseiros
Mais de 175	Dados muito grosseiros

Este índice é fácil de calcular, mas tem a limitação de medir só a preferência digital dos dígitos "0" e "5".

Índice de Whipple Modificado

Devido a esta limitação do índice original de Whipple, Spoorenberg (2017) propôs um índice modificado que contempla todos os dígitos finais e não só "0" e "5". O primeiro passo consiste em calcular índices específicos de atração para todos os dígitos, da seguinte forma proposta por Noumbissi (1992):

$$W_0 = 5 \cdot (P_{30} + P_{40} + P_{50} + P_{60}) \,/\, ({}_5P_{28} + {}_5P_{38} + {}_5P_{48} + {}_5P_{58}) \quad \text{(intervalo 28 – 62)}$$

$$W_1 = 5 \cdot (P_{31} + P_{41} + P_{51} + P_{61}) \,/\, ({}_5P_{29} + {}_5P_{39} + {}_5P_{49} + {}_5P_{59}) \quad \text{(intervalo 29 – 63)}$$

$$W_2 = 5 \cdot (P_{32} + P_{42} + P_{52} + P_{62}) \,/\, ({}_5P_{30} + {}_5P_{40} + {}_5P_{50} + {}_5P_{60}) \quad \text{(intervalo 30 – 64)}$$

$$W_3 = 5 \cdot (P_{23} + P_{33} + P_{43} + P_{53}) \,/\, ({}_5P_{21} + {}_5P_{31} + {}_5P_{41} + {}_5P_{51}) \quad \text{(intervalo 21 – 55)}$$

$$W_4 = 5 \cdot (P_{24} + P_{34} + P_{44} + P_{54}) \,/\, ({}_5P_{22} + {}_5P_{32} + {}_5P_{42} + {}_5P_{52}) \quad \text{(intervalo 22 – 56)}$$

$$W_5 = 5\cdot(P_{25} + P_{35} + P_{45} + P_{55}) / ({}_5P_{23} + {}_5P_{33} + {}_5P_{43} + {}_5P_{53}) \quad \text{(intervalo 23 – 57)}$$
$$W_6 = 5\cdot(P_{26} + P_{36} + P_{46} + P_{56}) / ({}_5P_{24} + {}_5P_{34} + {}_5P_{44} + {}_5P_{54}) \quad \text{(intervalo 24 – 58)}$$
$$W_7 = 5\cdot(P_{27} + P_{37} + P_{47} + P_{57}) / ({}_5P_{25} + {}_5P_{35} + {}_5P_{45} + {}_5P_{55}) \quad \text{(intervalo 25 – 59)}$$
$$W_8 = 5\cdot(P_{28} + P_{38} + P_{48} + P_{58}) / ({}_5P_{26} + {}_5P_{36} + {}_5P_{46} + {}_5P_{56}) \quad \text{(intervalo 26 – 60)}$$
$$W_9 = 5\cdot(P_{29} + P_{39} + P_{49} + P_{59}) / ({}_5P_{27} + {}_5P_{37} + {}_5P_{47} + {}_5P_{57}) \quad \text{(intervalo 27 – 61)} \qquad \text{(16.2.a-j)}$$

Nota-se que o denominador varia entre estes índices específicos e que cada denominador cobre 20 anos, em vez da totalidade de 40 anos como em (16.1). O índice total proposto por Spoorenberg agora consiste em somar os índices específicos da seguinte forma:

$$W_{tot.} = \sum_{i=0}^{9} |W_i - 1| \qquad (16.3)$$

Este índice, diferente do índice convencional de Whipple, tem um valor de 0 na ausência de qualquer atração digital e ele geralmente não é multiplicado por 100. O seu valor máximo é 18.

Índice Combinado de Myers

O índice de Myers (Myers, 1940; Naciones Unidas, 1955; Siegel e Swanson, 2004) que, igual ao anterior, considera a atração de todos os dígitos, calcula a proporção que representa a população que termina num determinado dígito em relação à população total, produzindo um índice de preferência para cada dígito final. Aqui o método é aplicado ao intervalo etário de 10-79 anos, mas isso nem sempre é aplicado da mesma forma e é possível encontrar aplicações com intervalos de 10-69 ou 20-79 anos. O "Combinado" ("Blended", em inglês) do método refere à forma como o método calcula a percentagem de dígitos finais. O mais simples seria calcular o percentual da população com idades no intervalo escolhido que terminam em "0", "1", "2" etc. Mas isso introduz um viés sistemático porque na maioria das populações, principalmente populações que ainda crescem rapidamente, os números diminuem com a idade, de modo que é de esperar que $P_{10} + P_{20} + P_{30} + P_{40} + P_{50} + P_{60} + P_{70}$ seja maior do que $P_{19} + P_{29} + P_{39} + P_{49} + P_{59} + P_{69} + P_{79}$. Para evitar este problema, o método é aplicado a 10 sequências consecutivas. Primeiro é aplicado ao intervalo de 10-69 anos, depois ao intervalo de 11-70 anos, e assim adiante, terminando com o intervalo de 19-78 anos. Depois todos os resultados são somados. Nesta soma, o número "10" aparece uma vez, "11" duas vezes etc. até "18" que aparece nove vezes. Todos os números entre "19" e "69" aparecem dez vezes, "70" aparece nove vezes, "71" oito vezes etc. até "78", que aparece uma só vez. Finalmente os resultados são organizados num esquema do seguinte tipo:

	10-19	20-29	30-39	etc.	60-69	70-79
0	Peso 1	Peso 10	Peso 10	....	Peso 10	Peso 9
1	Peso 2	Peso 10	Peso 10	....	Peso 10	Peso 8
2	Peso 3	....	....	....	....	Peso 7
3	Peso 4	....	....	....	....	Peso 6
4	Peso 5	....	....	....	....	Peso 5
5	Peso 6	....	....	....	....	Peso 4
6	Peso 7	....	....	....	....	Peso 3
7	Peso 8	....	....	....	....	Peso 2
8	Peso 9	....	....	....	....	Peso 1
9	Peso 10	....	....	....	....	Peso 0

Agora os números são somados linha por linha, com os seus devidos pesos. Por exemplo, no caso do dígito "0" a soma é $1 \cdot P_{10} + 10 \cdot P_{20} + 10 \cdot P_{30} + 10 \cdot P_{40} + 10 \cdot P_{50} + 10 \cdot P_{60} + 9 \cdot P_{70}$. Dividindo a soma de cada linha pelo total geral, se obtém a frequência relativa de cada dígito. As diferenças absolutas entre estas frequências relativas (em percentuais) e 10% são somadas e o resultado é dividido por 2, para eliminar a dupla contagem. O resultado é o índice de Myers que teoricamente pode variar de 0 a 90. Pode-se afirmar que inquéritos com declarações da idade que resulte em Índices de Myers próximos de, por exemplo, 10,0 poderiam se esforçar para obter melhores dados.

Além destes índices-padrão existem outros que possuem algumas vantagens teóricas em comparação com os mais comuns, tais como o índice de Bachi (1951), Ramachandran (1955), Carrier (1959) e Siegel (Siegel e Swanson, 2004), mas estes são pouco usados na prática. Vale mencionar que os módulos AGESEX e SINGAGE de PASEX (ver seção 17.2 do Capítulo 17) calculam vários dos índices mencionados nesta seção, bem como o índice das Nações Unidas abaixo.

Índice Combinado das Nações Unidas (Naciones Unidas, 1955)

Este índice é diferente dos demais na medida em que trata da estrutura etária global e não especificamente da preferência digital. A palavra "Combinado" neste caso não refere à soma de diferentes intervalos etários, como no caso do índice de Myers, mas à consideração da distribuição por sexo, além da distribuição por idades. O ponto de partida é a distribuição da população por sexo e idade, em intervalos quinquenais. Com esta informação calcula três indicadores preliminares:

$$RS_i = 100\ P_i^M\ /\ P_i^F \tag{16.4.a}$$

$$RI_i^M = 200\ P_i^M\ /\ (P_{i-1}^M + P_{i+1}^M) \tag{16.4.b}$$

$$RI_i^F = 200\ P_i^F\ /\ (P_{i-1}^F + P_{i+1}^F) \tag{16.4.c}$$

O primeiro indicador quantifica a relação de sexos por grupo etário. Os outros dois quantificam o desvio das populações masculina e feminina por grupo etário em relação aos grupos anterior e seguinte. Evidentemente este últimos indicadores não podem ser calculados para a primeira e a última faixa etária. O seguinte passo consiste em calcular a variação destes indicadores:

$$\Delta RS_i = RS_i - RS_{i-1} \tag{16.5.a}$$

$$\Delta RI_i^M = RI_i^M - 100 \tag{16.5.b}$$

$$\Delta RI_i^F = RI_i^F - 100 \tag{16.5.c}$$

De novo, estes indicadores não são definidos para a primeira e a última faixa etária. Finalmente o índice das Nações Unidas é definido com

$$INU = (3 \sum |\Delta RS_i| + \sum |\Delta RI_i^M| + \sum |\Delta RI_i^F|) \,/\, (k-2) \tag{16.6}$$

onde *k* é o número de faixas etárias quinquenais. Um valor menor de 20 é qualificado como "bom", entre 20 e 40 "mau" e mais de 40 como "muito mau".

Os índices explicados acima estão entre os instrumentos mais usados da análise demográfica. Andrade et al. (2016) fazem uma boa revisão destes Índices, dos esforços feitos para torná-los mais sensíveis e apresentam alguns resultados para América Latina. Os respectivos cálculos podem ser programados muito facilmente numa planilha de EXCEL, o qual contribui à sua popularidade. Entretanto, é preciso alertar para as suas limitações para efeitos da medição da qualidade dos censos, que são basicamente quatro:

1. Os primeiros três índices medem apenas a preferência digital e não captam as distorções na distribuição por idades que podem ocorrer como resultado do significado especial de determinadas idades específicas como 18 (maioria de idade), 30 (fim da juventude) ou 50 (em alguns países, a idade a partir da qual o recenseador não precisa mais preencher as perguntas sobre fecundidade das mulheres).

2. Nenhum dos índices capta certas outras distorções sistemáticas, como a tendência ao exagero das idades nas pessoas mais velhas.

3. Mais importante é que estes índices são usados muitas vezes para quantificar não só a qualidade da informação sobre idades, mas a qualidade da informação censitária em geral. Esta prática é muito questionável na medida em que a qualidade de muitas outras perguntas (por exemplo, sobre migração ou atividade econômica) depende de problemas cuja origem é muito diferente daquela que determina a preferência digital.

4. Mais especificamente, na medida em que o nível de educação da população melhora, a preferência digital diminui consideravelmente, mas o mesmo não acontece necessariamente com outros problemas de recolha de informação censitária.

A maioria destes Índices baseia-se no pressuposto da regularidade; isto é, a menos que existam fenômenos que bruscamente alterem o comportamento demográfico de uma população, sua composição por idade deveria reproduzir um perfil suave ou regular. Se ao longo de todo ano

nasce, morre, emigra ou imigra um determinado volume de pessoas, mesmo que este volume aumente ou diminua no tempo, não deveriam existir oscilações do tipo apresentado, por exemplo, pelos dados na Tabela 16.1 que servem de ilustração sobre a aplicação de dois dos índices mais usuais: Myers e Whipple.

Tabela 16.1: Quantificação do grau de atração por determinados dígitos (Índice de Myers) ou pelos dígitos 0 e 5 (Índice de Whipple) em Moçambique (1980 e 2007), Brasil (2010), Bangladesh (2011) e Timor-Leste Rural (2010)

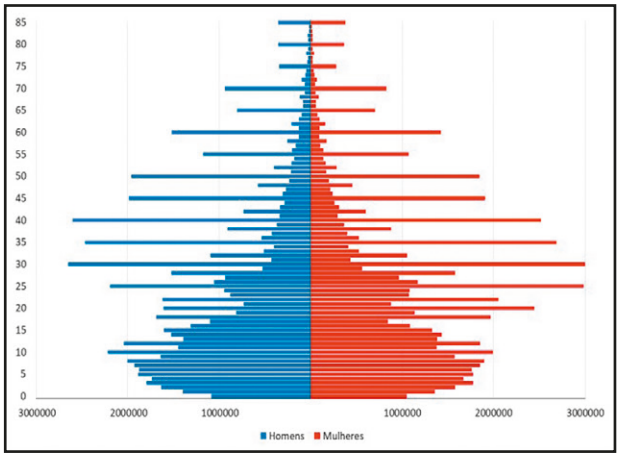
País	Data	Myers (10-79)		Whipple	
		Homens	Mulheres	Homens	Mulheres
Moçambique	1980	8,8	12,7	146,6	160,8
	2007	7,0	7,8	126,2	125,9
Brasil	2010	1,1	0,8	105,6	104,1
Bangladesh	2011	26,5	27,7	256,7	267,6
Timor-Leste Rural	2010	6,5	7,2	130,6	134,0

Fonte: Gaspar (1989); UNSD Demographic Statistics United Nations Statistics Division http://data.un.org/Data.aspx?-d=POP&f=tableCode%3A22; Processamento on-line do censo de Bangladesh com REDATAM; Timor-Leste: http://www.statistics.gov.tl/wp-content/uploads/2013/12/Publication_202_20 FINAL_20_20English_20Fina_Website.pdf.

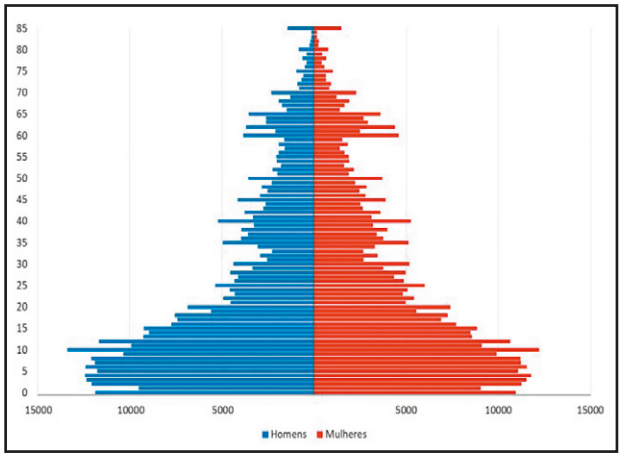
Apesar do que foi mencionado no ponto 4) acima, existem ainda hoje países cujos censos são afetados por uma preferência digital que qualifica os seus dados como "muito grosseiros". É o caso do Censo de 2011 de Bangladesh, cuja pirâmide etária se mostra no Gráfico 16.4 (esquerda). Segundo a Tabela 16.1 o índice de Whipple, tanto para homens como para mulheres, é maior de 175. Nenhum dos países de língua portuguesa em 2010 foi caracterizado por problemas de atração digital desta gravidade embora tanto os dados de Moçambique (2007) no Gráfico 16.2.B como os da área rural de Timor-Leste (2010) no Gráfico 16.4 (direita) ainda mostrem um padrão caracterizado por oscilações significativas de uma idade simples para outra. Segundo a Tabela 16.1, ambos os países se encontram na transição entre "dados aproximados" e "dados grosseiros". Nota-se que a pirâmide da direita do Gráfico 16.4 também parece indicar duas irregularidades reais. Há certa falta de população por volta dos 35 anos, que seria a coorte de nascimentos de 1975, o ano da invasão de Timor-Leste pela Indonésia que causou mais de 100.000 mortes. A outra particularidade notável é a maior população de 60-64 anos, a coorte nascida imediatamente depois da Segunda Guerra Mundial. Como ambas as particularidades provavelmente são reais, elas precisam ser tratadas com cuidado no momento de graduar/suavizar a distribuição etária para eliminar irregularidades espúrias.

Gráfico 16.4: Distribuição etária por idade simples para Bangladesh (Censo de 2011) e a área rural de Timor-Leste (2010)

Bangladesh (2011)



Timor-Leste Rural (2010)



Fontes: Processamento on-line do Censo de Bangladesh com REDATAM; http://www.statistics.gov.tl/wp-content/uploads/2013/12/Publication_202_20FINAL_20_20English_20Fina_Website.pdf.

A Tabela 16.2 apresenta os resultados do cálculo do Índice Combinado das Nações Unidas para o Censo de 2009 da Guiné-Bissau. A distribuição da população por idade e sexo claramente apresenta algumas irregularidades como a subida inexplicável da Razão de Sexos na faixa etária de 55-59 anos e a diminuição brusca da população de ambos os sexos por volta dos 40 anos.

Tabela 16.2: Cálculo do Índice Combinado das Nações Unidas para o Censo da Guiné-Bissau de 2009

Idades	Homens	Mulheres	RS(i)	|ΔRS(i)|	|ΔRI^{M}(i)|	|ΔRI^{F}(i)|
0-4	115,009	113,988	100,896			
5-9	104,650	103,357	101,251	0,355	2,186	1,945
10-14	89,814	88,782	101,162	0,089	4,704	6,856
15-19	83,844	87,276	96,068	5,095	5,838	5,448
20-24	68,624	76,751	89,411	6,656	3,561	1,558
25-29	58,472	68,656	85,167	4,245	7,925	12,774
30-34	39,733	45,008	88,280	3,113	14,166	16,858
35-39	34,109	39,612	86,108	2,172	5,961	8,924
40-44	24,647	27,725	88,898	2,790	13,158	14,796
45-49	22,654	25,467	88,954	0,056	12,338	11,409
50-54	15,685	17,993	87,173	1,782	12,274	7,154
55-59	13,105	13,292	98,593	11,420	2,108	10,365
60-64	9,984	11,665	85,589	13,004	2,371	7,151
65-69	7,348	8,481	86,641	1,051	1,929	4,061
70-74	5,001	6,015	83,142	3,499	7,833	6,925
75-79	3,504	4,444	78,848	4,294	27,872	29,191
80-84	4,715	6,537	72,128			
Total				59,621	124,222	145,415

Fonte: UNSD Demographic Statistics United Nations Statistics Division (2009) http://data.un.org/Data.aspx?d=PO-P&f=tableCode%3A22.

O valor combinado das três componentes do índice é calculado como

$$INU = (3 \cdot 59{,}621 + 124{,}222 + 145{,}415) / 15 = 65{,}85 \quad (16.7)$$

Este resultado sugere problemas consideráveis na enumeração da estrutura por idade e sexo. Entretanto, é sempre aconselhável analisar se existe alguma explicação plausível para as irregularidades observadas. Por exemplo, a irregularidade na estrutura etária por volta dos 40 anos poderia estar relacionada com a forte emigração que houve na década de 70. Caso essa explicação possa ser corroborada, o valor de (16.7) é irrelevante.

16.3.4 Métodos de ajuste e graduação

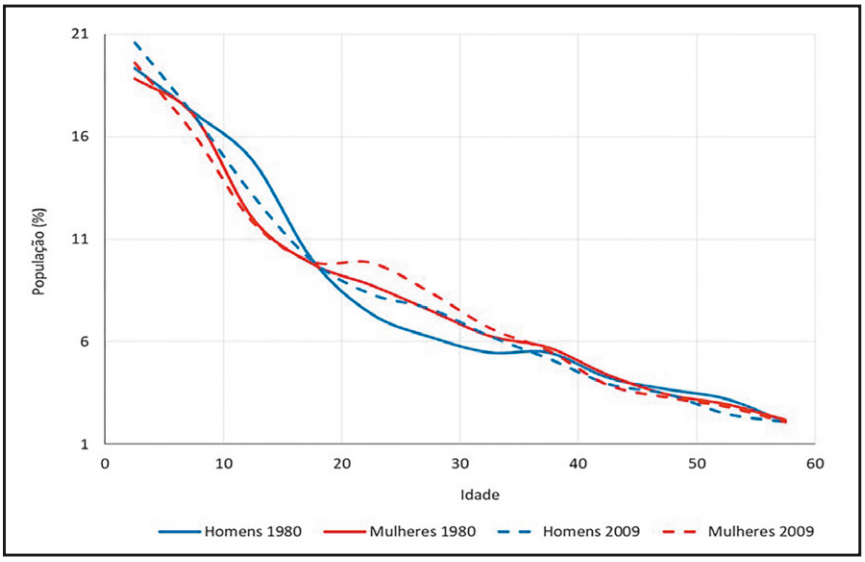
Como já foi assinalado algumas vezes anteriormente, é preciso ter cuidado no ajuste de dados, para não eliminar certas irregularidades aparentes que têm sua base em tendências reais. Há de se ater ao fato de que alterar a idade de um indivíduo, mudando portanto a composição por idade original, traz consequências nas análises que se farão com relação a outras características, seja que envolvam ou não o critério idade, seja que se façam utilizando métodos indiretos de análise. Um par de exemplos simples serve para mostrar a complexidade de ajustar ou corrigir uma distribuição da idade:

- Uma mulher erra na declaração da idade ou tende a arredondá-la, pode, ou não errar/arredondar a idade ao casar, ao nascimento do filho, ao primeiro uso de contracepção etc. Ao ajustar apenas a declaração da idade da mulher altera-se a relação entre a idade e as outras variáveis.

- Se a população no geral, tem tendência a declarar a idade em determinado dígito, pode manter (ou não) essa tendência ao declarar a idade dos membros que saíram do domicílio (agregado familiar); dos anos transcorridos desde que saíram do mesmo; dos anos que residiu em outro lugar; da idade dos que morreram em determinado período. Da mesma que no caso anterior, ao ajustar apenas a declaração da idade da população, altera-se a relação entre esta e as outras respostas dadas pelo informante.

Como se pode deduzir destes exemplos, alterar a distribuição por idade de uma população pode causar vieses em relação ao cálculo de outros indicadores. É por esta razão que se opta, muito frequentemente, por considerar a declaração da idade em grupos quinquenais e não ajustar (ou corrigir) a distribuição original por idades simples. É consenso que este nível de agregação permite superar em grande parte o efeito de atrações por dígitos que –como visto– trata-se de dígitos vizinhos, sem perder o poder de discriminação desta característica. O Gráfico 16.5 mostra o comportamento bastante diferenciado por idade da população moçambicana mesmo que agrupada quinquenalmente. Os desvios causados pelas preferências de determinados dígitos são obviados, mas outras características se mantêm: os anos 80 apresentam uma relativa falta de população masculina nas idades jovens adultas que clama por explicação, seja esta a omissão, conflitos sociais da época etc.

Gráfico 16.5: Moçambique, 1980 e 2009 - Distribuição relativa da população por grupos de idade quinquenal (homens e mulheres)



Fontes: Gaspar (1989); UNSD Demographic Statistics- United Nations Statistics Division http://data.un.org/Data.aspx?d=POP&f=tableCode%3A22.

Chackiel e Macció (1979) fizeram algumas propostas para otimizar o procedimento de agregação em intervalos quinquenais. A ideia é que os agrupamentos convencionais em *(x, x+4)* e *(x+5, x+9)*, onde *x* é uma dezena, implicam que cada intervalo quinquenal começa com uma idade sujeita a uma forte atração digital que pode "roubar" população do intervalo anterior. Para evitar o problema, eles propuseram usar intervalos não convencionais do seguinte tipo:

Primeira alternativa: *(x+3, x+7)* e *(x+8, x+12)*
Segunda alternativa: *(x+0,5, x+4,5)* e *(x+5,5, x+9,5)*

Na primeira alternativa, as idades que atraem população ficam no meio dos intervalos, não no extremo, de modo que a grande maioria dos erros de declaração provavelmente ocorreria só *dentro* do intervalo. Na segunda alternativa, a população que corresponde às idades de atração seria dividida igualmente entre os dois intervalos. Em casos onde a atração digital é suficientemente forte para afetar a distribuição da população entre intervalos quinquenais, é de esperar que essas divisões sejam menos sensíveis a erros do que a convencional.

Resta o problema, entretanto, de como converter os intervalos não convencionais assim construídos em intervalos convencionais. Isso exige algum tipo de interpolação. As técnicas apropriadas para tal fim serão discutidas na seção 17.5 do Capítulo 17. O outro problema que pode surgir é que, além de "0" e "5", pode haver outros dígitos que atraem população, como "2" ou "8". Se for assim, pelo menos a primeira alternativa não é recomendável. Mas podem existir outras alternativas, que podem ser identificadas mediante a análise da atração de cada dígito (ver fórmulas 16.2.a-j ou seu equivalente no caso do índice de Myers), e que sim, resultam numa divisão equilibrada.

16.3.4.1 Avaliação da omissão por idade ou sub-registro

Um outro aspecto que se deve considerar ao avaliar a informação por idade, ademais do erro, é a omissão, entendendo esta como a ausência de registro ou subenumeração. Todas as idades ou grupos de idades devem ser objeto de avaliação, sendo que as crianças menores de cinco (ou dez) anos devem ser consideradas com especial ênfase dada sua importância demográfica. Elas são resultado combinado da fecundidade, da mortalidade na infância e da migração dos 5 (ou 10) anos anteriores à data do inquérito. Portanto deve existir coerência entre as medidas destas variáveis e a contagem dessas crianças (United Nations, 2017 b).

Com relação às crianças, como dito, a tendência a omitir os menores de 5 (ou 10) anos num inquérito seja censitário ou amostral é uma constatação histórica e praticamente universal e está relacionado tanto a atitudes socioculturais como conjunturais e até circunstanciais. No primeiro caso, por exemplo, a criança pode não ser considerada "uma pessoa", principalmente se é menor de um ano, um mês ou uma semana, e daí ser omitida ao fazer o inventário das "pessoas" residentes. Casos de informantes alheios ao núcleo familiar ou domiciliar, podem igualmente, não notar a existência de crianças, novamente, principalmente, das mais novas. Nos casos de omissão territorial geral por deficiências estruturais, se as crianças costumam ser maiores em número, o que é o caso de países em desenvolvimento, a sua omissão seria proporcionalmente maior. Em razão das considerações anteriores, é altamente recomendável avaliar o contingente declarado de crianças comparando-o com outras fontes e/ou outros períodos. São várias as alternativas à disposição para efetuar esta avaliação e todas elas seguem a lógica da *Equação de Equilíbrio Demográfico* ou *Compensadora*.

1. Com os dados da mesma fonte, seja censo ou inquérito amostral: estimando taxas de fecundidade por idade para o quinquênio (ou decênio) anterior e aplicando-as à população feminina é possível replicar o número de nascimentos ocorridos nesse quinquênio (ou decênio) e, mediante estimativas de mortalidade e migração obter o número de sobreviventes residentes menores de 5 (ou 10) anos que deviam ter sido declarados.

2. Se há disponibilidade de inquérito posterior: mediante uma retroprojeção da população menor de 5 (ou 10 anos) a partir do segundo inquérito é possível estimar o número prévio de residentes que deviam ter sido registrados no primeiro inquérito.

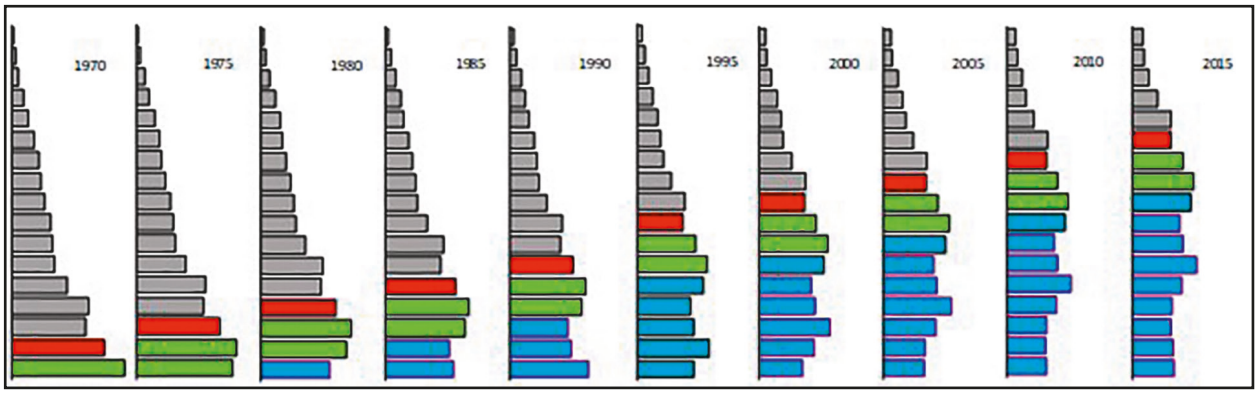
3. Se há disponibilidade de registros vitais de cobertura relativamente completa: é possível, calcular a população menor de 5 (ou 10) anos a partir dos registros de nascidos vivos do período anterior correspondente como passo inicial. Subtrai-se, logo, o número de óbitos pertencentes a estes nascimentos, ocorridos até o momento do levantamento censitário ou inquérito e incorpora-se o eventual saldo migratório também, ocorrido no quinquênio (ou decênio) anterior.

As duas primeiras alternativas requerem informação demográfica adicional que frequentemente pode obter-se indiretamente via aplicação de técnicas indiretas (ver Moultrie et al., 2013; Preston, Heuveline e Guillot, 2001). A terceira alternativa é um exemplo do cálculo direto de uma *Equação de Equilíbrio Demográfico* (ver Capítulo 7).

Com relação aos demais grupos etários, uma forma direta de avaliar a declaração por idade quando ela existe para mais de um momento no tempo é acompanhar as coortes no período definido por esses momentos e estimar uma razão de sobrevivência dessas coortes. Esta razão deve corresponder aos níveis de mortalidade estabelecidos para essa população nesse período pressupondo que não existe migração e a cobertura dos dois inquéritos é a mesma.

Assim, se fortes oscilações na declaração da idade entre várias coortes ou gerações num primeiro momento, são reais, assumindo que se trata de uma população que, no período não recebeu ou originou fluxos migratórios e que a mortalidade não mudou bruscamente, tais oscilações deverão se manter, entre as mesmas coortes, nos levantamentos posteriores. Um exemplo clássico é dado pelas bruscas oscilações na distribuição por idade de países expostos no passado a violentas guerras, como a França ou Japão no passado ou a intervenções políticas como o caso da China, ilustrado no Gráfico 16.6 com a distribuição estilizada por grupos quinquenais da população total apresentada nas publicações das Nações Unidas, isto é, já ajustadas e/ou corrigidas.

Gráfico 16.6: China, 1970-2015: População por idade quinquenal - distribuição relativa



Fonte: Divisão de População das Nações Unidas, Revisão de 2019.

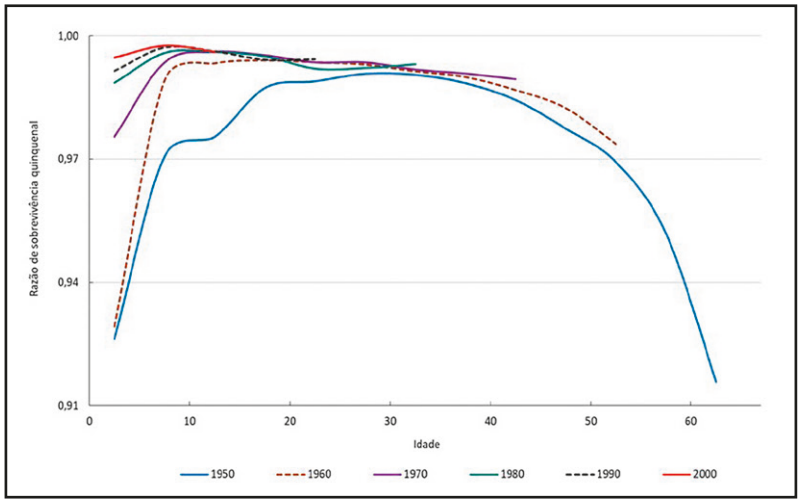
A primeira série (1970) apresenta, para as idades centrais uma relativamente menor quantidade de população em razão das perdas com a política do "Salto para Frente" dos anos 1958-1962 e que se observa, ainda, nos períodos seguintes. Nos anos 80, o menor tamanho no grupo etário mais novo (0-4), é a consequência da política do filho único e também, do menor tamanho das coortes que nesse momento estão nas idades mais reprodutivas; essa coorte ou geração de idades 0-4 aparece também, nos anos seguintes, sendo que em 2015 já está com idades 35-39. Em 1990, em que pese o vigor da política do filho único, há um aumento no tamanho da mais nova geração que surge nesse ano e que se deve, novamente ao volume da população em idade reprodutiva, que desta vez, é volumosa e que nasceu antes da implantação da tal política. Assim, a China é um exemplo de que nem sempre, oscilações na distribuição da idade são erros de declaração ou inconsistências.

Mesmo na presença de oscilações no padrão por idade nos diferentes momentos, as razões de sobrevivência, introduzidas em (9.15) do Capítulo 9, utilizando um olhar longitudinal mostram uma extinção das gerações, paulatina e consistente, como se pode verificar no Gráfico 16.7. Efetivamente, a comparação em períodos quinquenais consecutivos do volume de uma mesma geração num dado grupo etário nunca é superior a 1,0 assinalando que nunca uma coorte

aumenta seu contingente original. Dadas suas dimensões continentais e sua organização social, pode-se assumir que a China é um pais fechado a movimentos migratórios capazes de alterar as correlações intergeracionais.

Todavia, na medida em que o tempo passa, as gerações nascidas de 1950 em diante, contempladas, antes, nas pirâmides do Gráfico 16.7 diminuem, proporcionalmente, cada vez menos, de acordo aos diversos regimes de mortalidade em declínio que a China tem experimentado. A geração nascida em 1950 mostra, para todas as idades, as menores razões de sobrevivência, sinalizando a relativa maior mortalidade a que esta geração sempre se submeteu; mesmo à idade 60, atingida no século XXI, é ela quem apresenta a menor razão de sobrevivência se comparada com as gerações mais novas. As irregularidades notadas anteriormente na distribuição por idade são coerentes com as irregularidades que as razões de sobrevivência apresentam: as gerações dos anos 50 e 60 foram as mais atingidas pelas consequências do "Salto para Frente".

Gráfico 16.7: China, 1950 - 2010: Razões de Sobrevivência quinquenais para gerações nascidas nos anos indicados



Fonte: Divisão de População das Nações Unidas, Revisão de 2015.

Para os casos onde há necessidade de corrigir uma distribuição etária quinquenal irregular, existe um bom número de propostas baseadas, seja em evidências concretas e interpolações matemáticas a partir de modelos de população teóricas, como em modelagens estatísticas com algum grau de complexidade computacional. Algumas das alternativas disponíveis serão discutidas a seguir. Desde já é importante alertar que essas técnicas podem ajudar a corrigir transferências indevidas de população entre diferentes categorias etárias, mas elas não corrigem a subenumeração. Se as irregularidades na estrutura etária se devem à subenumeração diferencial por idades, elas não deveriam ser usadas pois o seu resultado seria apenas a redistribuição do erro entre diferentes categorias.

16.3.4.2 Médias móveis

O conceito de médias móveis ponderadas já foi introduzido em (9.35). No caso em que os erros se alternam (um intervalo superdeclarado com um erro e, o próximo subdeclarado com um

erro igual, e assim adiante), Chackiel e Macció sugerem que a seguinte média móvel (originalmente introduzida por United Nations, 1956) dá bons resultados:

$$\overline{{}_5P_x} = (-\,{}_5P_{x-10} + 4\,{}_5P_{x-5} + 10\,{}_5P_x + 4\,{}_5P_{x+5} - {}_5P_{x+10})/16 \qquad (16.8)$$

Entretanto, é preciso observar que o suposto de erros alternantes com tamanhos iguais pode não ser realista. Sob o suposto mais realista de erros relativos iguais e alternantes, (16.8) pode ser substituída pela seguinte alternativa:

$$\overline{{}_5P_x} = \exp\left((-\ln({}_5P_{x-10}) + 4\ln({}_5P_{x-5}) + 10\ln({}_5P_x) + 4\ln({}_5P_{x+5}) - \ln({}_5P_{x+10}))/16\right) \qquad (16.9)$$

Fórmulas de Arriaga

Arriaga (1968) também desenvolveu duas fórmulas conhecidas como as fórmulas de suavização branda e suavização forte. A expressão para a suavização branda é a seguinte:

$$\overline{{}_5P_{x+5}} = (-\,{}_{10}P_{x-10} + 11\,{}_{10}P_x + 2\,{}_{10}P_{x+10})/24 \quad \text{e} \quad \overline{{}_5P_x} = {}_{10}P_x - \overline{{}_5P_{x+5}} \qquad (16.10)$$

Nota-se que (16.10), diferentemente de (16.8), funciona com base em intervalos decenais e os desagrega em intervalos quinquenais. Normalmente *x* é uma idade que termina em 0. Se *x=0*, ou seja no grupo etário mais baixo, se usa a seguinte expressão:

$$\overline{{}_5P_{x+5}} = (8\,{}_{10}P_x + 5\,{}_{10}P_{x+10} - {}_{10}P_{x+20})/24 \quad \text{e} \quad \overline{{}_5P_x} = {}_{10}P_x - \overline{{}_5P_{x+5}} \qquad (16.11)$$

No caso do último grupo etário, o mesmo procedimento é invertido:

$$\overline{{}_5P_x} = (-\,{}_{10}P_{x-20} + 5\,{}_{10}P_{x-10} + 8\,{}_{10}P_x)/24 \quad \text{e} \quad \overline{{}_5P_{x+5}} = {}_{10}P_x - \overline{{}_5P_x} \qquad (16.12)$$

A suavização forte consiste em suavizar primeiro as populações por intervalo decenal

$$\overline{{}_{10}P_x} = ({}_{10}P_{x-10} + 2\,{}_{10}P_x + {}_{10}P_x)/4 \qquad (16.13)$$

antes de usar (16.10-12) para desmembrar os intervalos quinquenais.

16.3.5 Método de Carrier e Farrag

No caso onde as distorções na estrutura quinquenal por idades consiste em transferências de população entre o grupo de 0-4 e 5-9, 10-14 e 15-19, 20-24 e 26-29 etc., mas onde os totais

por grupo decenal podem ser considerados aproximadamente corretos, a alternativa sugerida por Carrier e Farrag (1961) consiste em aplicar a seguinte correção:

$$\overline{{}_5P_{10}} = {}_{10}P_{10}/2 + ({}_{10}P_0 - {}_{10}P_{20}) / 16 \quad (16.14.a)$$

$$\overline{{}_5P_{15}} = {}_{10}P_{10}/2 - ({}_{10}P_0 - {}_{10}P_{20}) /16 \quad (16.14.b)$$

e assim adiante para os outros grupos decenais, exceto os extremos.

16.3.6 Método das ogivas oblíquas

Este método, na descrição de Chackiel e Macció (1979), pode ser usado para ajustar dados etários quinquenais que sofrem de distorções graves, mas tem o inconveniente de ser relativamente arbitrário. Além disso, como sempre, não deve ser usado em situações onde as distorções podem ser o resultado de tendências reais. O primeiro passo consiste em acumular os dados etários quinquenais até um limite superior conveniente. No caso da Tabela 16.3 o limite escolhido foi 65 anos, mas também poderia ser 70, 75 ou 80. O segundo passo consiste em subtrair dessa função acumulada a função que seria observada se todos os grupos etários fossem do mesmo tamanho. Ou seja, se SP_x descreve a população acumulada até a idade de x anos, o resultado seria

$$O_x = SP_x - SP_{65} \cdot x / 65 \quad (16.15)$$

Esta função O_x é chamada a *ogiva oblíqua* e ela aparece na coluna D da Tabela 16.3. Tipicamente é uma função côncava mais ou menos regular que começa e termina em 0. Chackiel e Macció recomendam manualmente traçar um gráfico suave que regularize a forma definida pelos pontos assim definidos. Entretanto, esse é um procedimento bastante subjetivo que não garante a correspondência mais próxima entre os pontos observados e a curva de ajuste. Um procedimento mais objetivo consiste em reconhecer que a curva em muitos casos poderá ser representada pela seguinte fórmula:

$$\hat{O}_x = C\,(x/65)^{\alpha}\,(1-x/65)^{\beta} \quad (16.16)$$

onde C é um fator de nível e α e β determinam a forma da curva (mais inclinada para a esquerda, mais inclinada para a direita, mais larga ou mais pontuda). Para determinar os valores ótimos de α, β e C, é preciso usar o recurso do Solver em EXCEL[1]. Inicialmente, coloque valores arbitrários de α, β e C em B17, B18 e B19 (por exemplo B17=1, B18=1, B19=1000000). Com esses valores arbitrários, calcule $\hat{O}_x$ usando (16.16) e coloque-o na coluna E. Na coluna F, calcule a diferença quadrada entre O_x e $\hat{O}_x$, ou seja, $(O_x - \hat{O}_x)^2$. A fórmula para a soma de todos os desvios da coluna F (Desvio

[1] Se nunca usou o Solver antes, provavelmente não está instalado na sua versão de EXCEL, mas a instalação é simples seguindo os passos indicados na facilidade de Ajuda do EXCEL.

Total) deve ser colocada em B20. Agora ative o Solver e entre com as instruções para minimizar B20 por meio da variação de B17, B18 e B19. O resultado que aparecerá são os valores de B17, B18, B19 e B20 (e das colunas E e F) que estão na Tabela 16.3. Com esses valores otimizados de α, β e C agora é possível calcular as populações ajustadas da coluna G pela fórmula

$$\overline{{}_5P_x} = \hat{O}_{x+5} - \hat{O}_x + 5 \cdot SP_{65} / 65 \qquad (16.17)$$

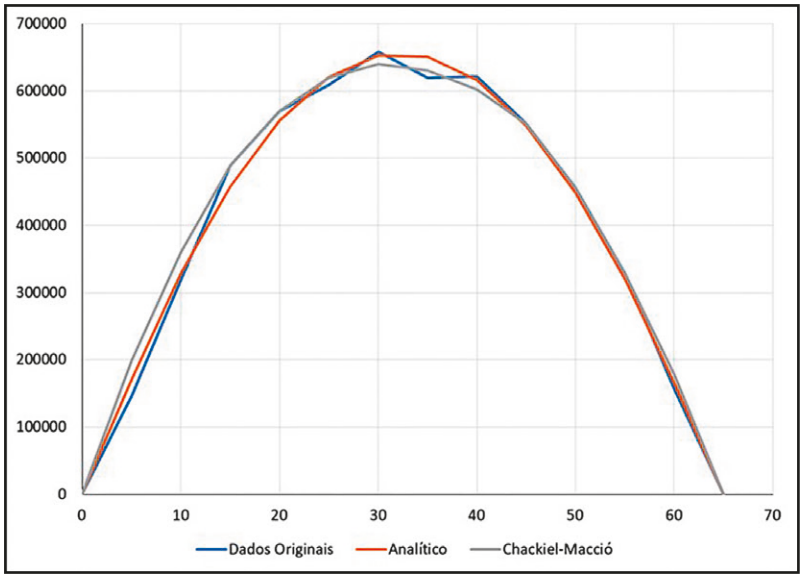
Esses valores são comparados com os obtidos por Chackiel e Macció na coluna H que não se baseiam numa fórmula explícita, mas na suavização manual dos dados. Como se pode apreciar pela comparação de B20 com H20, o Desvio Total dos valores sugeridos por Chackiel e Macció é quase o dobro do obtido pelo procedimento sugerido aqui. Portanto, o procedimento explicado acima fornece um ajuste mais próximo, além da vantagem de ser mais explícito e portanto menos arbitrário do que o procedimento gráfico sugerido no artigo original que foi escrito antes da existência de recursos como EXCEL. O Gráfico 16.8 mostra as três ogivas, para fins de comparação.

Tabela 16.3: Ilustração do método da ogiva oblíqua com os dados etários do Censo do Haiti em 1950, comparando a variante analítica com a gráfica proposta por Chackiel e Macció

	A	B	C	D	E	F	G	H
1	Idades	População	Acumulada	Ogiva	Ajustada	Desvios	População	Chackiel-Macció
2	0-4	374872	0	0	0	0	399807	427908
3	5-9	400518	374872	146964	171899	621763625	384776	387909
4	10-14	397708	775390	319573	328766	84516140	356889	357281
5	15-19	308026	1173098	489373	457747	1000196459	325770	308026
6	20-24	267401	1481124	569490	555608	192716605	293152	278418
7	25-29	277177	1748525	608983	620852	140865173	259817	247908
8	30-34	189144	2025702	658252	652761	30151025	226258	217909
9	35-39	229644	2214846	619487	651111	1000046071	192877	200408
10	40-44	157697	2444490	621223	616080	26454337	160089	176420
11	45-49	133451	2602187	551012	548260	7569639	128436	133451
12	50-54	99389	2735638	456554	448788	60314070	98813	99389
13	55-59	56828	2835027	328035	319693	69589793	73243	77874
14	60-64	70954	2891855	156954	165027	65170386	62881	49908
15	65+		2962809	0	0	0		
16								
17	Alfa	1,07215						
18	Beta	1,08857						
19	Nível	2933807						
20	Desvio T.	3299353323					Desvio T.	5805677432

Fonte: Cálculos baseados em Chackiel e Macció (1979): Cuadro 4.

Gráfico 16.8: Representação gráfica das três ogivas calculadas na Tabela 16.3

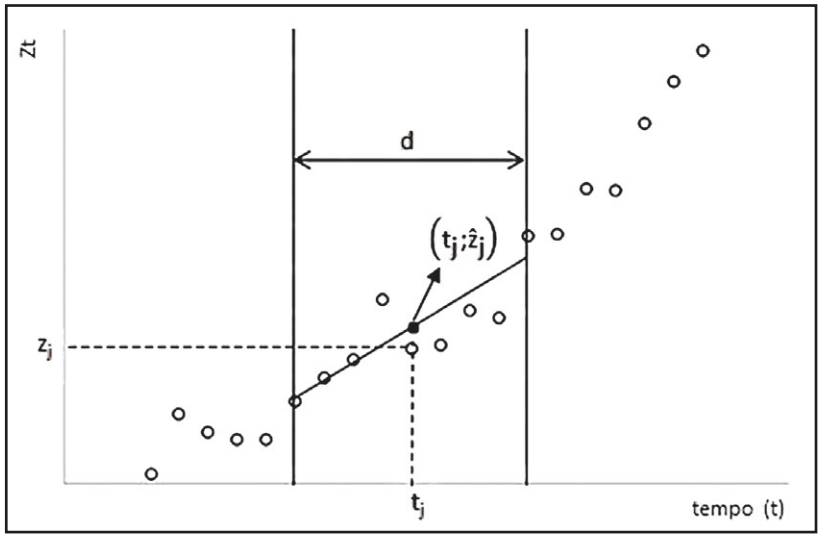


Fonte: Elaboração própria com base em Chackiel e Macció (1979).

16.3.7 O método LOWESS de regressão local

O LOWESS (Locally Weighted Regression Scatter Plot Smoothing) é um método bastante utilizado para suavização de dados (números, taxas, proporções etc.), especialmente quando esses dados estão expostos ao longo do tempo. Trata-se de um método de suavização local por utilizar um número relativamente pequeno de pontos ao redor do ponto no qual se deseja suavizar o dado. Também é ponderado pelo fato de utilizar um método de estimação conhecido por Mínimos Quadrados Ponderados, muito útil para reduzir a influência de valores discrepantes nos dados ao se produzir uma estimativa. A Figura 16.1 ajuda a entender a dinâmica do método LOWESS. Supõe-se uma série de dados Z_t representada pelos pares de pontos $(t_j;)$, onde t_j é o instante em que , o valor suavizado de z_j, é observado. Considere um intervalo de d pontos igualmente espaçados ao redor do ponto $(t_j;)$. Na Figura 16.1, há $d = 9$ pontos e está delimitado por duas faixas contínuas verticais. No método LOWESS, os $d = 9$ pontos são utilizados para suavizar o dado z_j no instante t_j, que resulta no ponto de partes ordenados $(t_j;)$. O número de pontos d é definido fazendo $d = (p_N)$, onde p é a proporção de pontos a serem utilizados $(0 < p < 1)$. Quanto maior o valor de p mais suave será o ajustamento da série. No limite, quando $p = 0$, os valores ajustados (suavizados) se igualam aos valores observados.

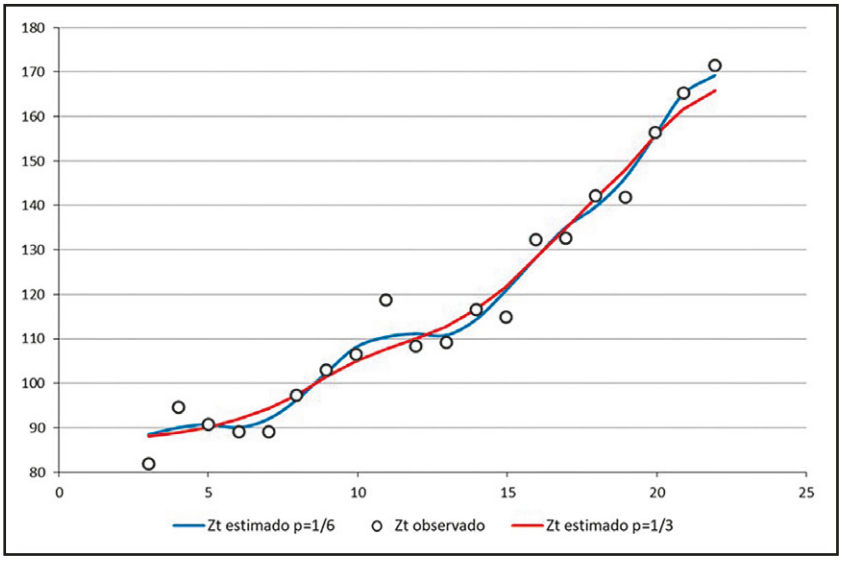
Figura 16.1: Representação esquemática do processo LOWESS



Fonte: Adaptado de Morettin e Toloi (2006).

Intuitivamente, o que o método LOWESS faz é definir pesos para os pontos vizinhos de $(t_j;)$, dentro da faixa de pontos q, de forma que, para cada peso definido, os pontos vizinhos tenham pesos decrescentes à media que o dado observado Z_t se afasta de t_j. Definidos os pesos, ajusta-se uma reta ao número de pontos q pelo Método de Mínimos Quadrados. O Gráfico 16.9 apresenta uma aplicação do método LOWESS aos dados Z_t, apresentados na Figura 16.1, considerando valores de p iguais a 1/3 e 1/6. Na curva suavizada azul o valor de p é mais próximo de zero e, portanto, produz uma suavização menor.

Gráfico 16.9: Ilustração gráfica da aplicação do processo LOWESS



Fonte: Adaptado de Morettin e Toloi (2006).

Esta aplicação do método LOWESS foi feita no software "R", que será introduzido mais sistematicamente no próximo capítulo, com as seguintes linhas de comando:

```
zt.est1 = lowess(zt, f = 1/3)
zt.est2 = lowess(zt, f = 1/6)
```

Nesta linha de comandos, "zt" é a série observada, *f = 1/3* e *f = 1/6* definem os graus de suavização desejados, "lowess" é a função interna do "R" para a aplicação do método e "zt.est1" e "zt.est2" são as séries suavizadas. Como esta função aplica o método automaticamente, o usuário não precisa programar nada.

16.4 A DECLARAÇÃO DA IDADE E A COBERTURA CENSITÁRIA

A comparação de uma mesma coorte em dois momentos, usando a lógica explicada nas seções 7.3 e 7.4 do Capítulo 7, oferece uma primeira aproximação mais quantitativa de eventuais omissões na declaração da idade. Para tal propósito se usam as razões de sobrevivência de uma tábua de vida aproximada. O mesmo procedimento também é usado para avaliar a cobertura da população por idade feita por dois ou mais inquéritos e a presença de movimentos migratórios (Carvalho, 1996).

A Tabela 16.4 ilustra como as razões de sobrevivência intercensitárias (RIS) permitem avaliar a omissão por idade da população. Mostra a população brasileira menor de 35 anos, recenseada em 2000, e o mesmo grupo etário – ou gerações – retratada 10 anos mais tarde, no Censo de 2010 e que corresponderia à população de 10-44 anos completos. Assumindo que se trata de uma população fechada em ambos censos e que os erros de declaração da idade são mínimos, a comparação dos totais indica que esta coorte teria perdido 76,37 mil indivíduos da coorte inicial composta de 111,26 milhões ou –mediante o cálculo da RIS– que 99,93% teria sobrevivido até a idade 45. Isto equivale a um nível de mortalidade excepcionalmente baixo, inexistente na atualidade e não esperado no futuro próximo; esta primeira inconsistência indica que houve sub-registro no primeiro recenseamento ou sobrerregistro no segundo.

Análises mais detalhadas podem ser feitas considerando as coortes formadas pelos diversos grupos etários. A coorte mais jovem (0-4 anos), composta, segundo a Tabela 16.2, por aproximadamente 16,4 milhões de crianças em 2000, é recenseada 10 anos depois e resulta num volume de 17,2 milhões (ou RIS superior a 1,0). Se, como registra a literatura clássica sobre a qualidade dos censos, a população de idades 10-14 anos e ainda 15-19 anos corre, costumeiramente, muito menos risco de ser omitida, os dados indicam que houve sub-registro dos menores de 5 anos em 2000. Esta questão é analisada em mais detalhe por Santos e Gonçalves (2018). O mesmo pode-se afirmar do grupo de 5-9 anos. A explicação alternativa, embora menos plausível porque raramente constatada seria a de sobre-enumeração em 2010 da população de 10-19 anos.

Tabela 16.4: Brasil, 2000 e 2010. População segundo grupos etários quinquenais de menos de 30 anos em 2000 e de 10-39 anos em 2010, tal como declarados nos correspondentes censos (em milhares) e razão de sobreviventes

População por idade em 2000		População por idade em 2010		Razão Intercensitária de Sobrevivência (RIS) em 2010
0-4	16.376			
5-9	16.542			
10-14	17.348	10-14	17.167	1,0483
15-19	17.940	15-19	16.991	1,0271
20-24	16.142	20-24	17.245	0,9941
25-29	13.850	25-29	17.104	0,9534
30-34	13.029	30-34	15.745	0,9754
		35-39	13.889	1,0028
		40-44	13.009	0,9985
Total	111.226	Total	111.150	0,9993

Fonte: IBGE - Censos Demográficos de 2000 e 2010.

Já os três seguintes grupos etários (de 10-24 anos em 2000) mostram uma diminuição no volume da coorte registrada em 2010. A RIS destes grupos etários, no entanto, mostra uma tendência errática. De acordo ao que se sabe da mortalidade, depois da idade 10, os riscos de morte aumentam com a idade, opostamente ao que se observa na série brasileira. Todavia, a coorte de idades 25-29, além de registrar novamente um aumento com relação à coorte anterior, esse aumento torna a RIS superior a 1,0.

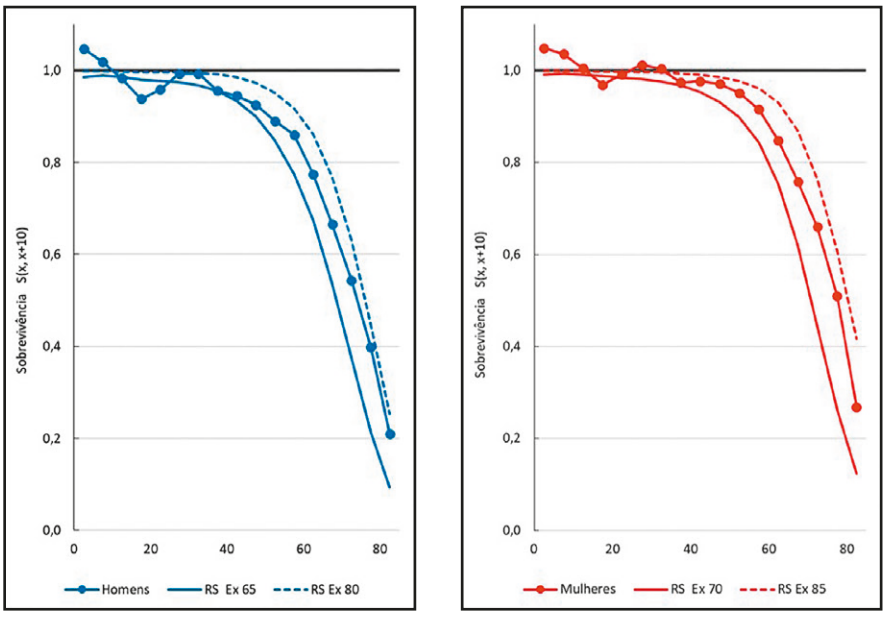
O comportamento da RIS, atendendo aos valores obtidos para cada grupo etário, indica em primeiro lugar que, como no caso da população mais jovem, houve sub-registro no primeiro recenseamento ou sobrerregistro no segundo. Esta afirmação descansa na comparação das RIS por idade com níveis de mortalidade estimados para estas populações. A probabilidade de sobrevivência nestes grupos etários das tábuas de mortalidade brasileiras para estes períodos é sempre menor que as RIS calculadas na Tabela 16.4. Em segundo lugar, a tendência errática indica que os erros de cobertura teriam sido diferenciais por grupos etários.

É importante assinalar que o fato de ter RIS extremamente alta tanto entre crianças (0-10 anos) como entre a população em idades centrais do ciclo reprodutivo (20-24 e 25-29) sugere reavaliar o pressuposto de se tratar de uma população fechada. Se houve uma entrada ao país de população em idade produtiva –e reprodutiva– no período intercensitário, com o consequente impacto na RIS destas idades, a RIS alterar-se-ia, assim como a das idades mais jovens. Por ser o Brasil um país de dimensões colossais, a imigração teria que ter sido maciça e que aparentemente não ocorreu. Com tudo, nesta hipotética situação, a afirmação de sub e sobrerregistro ficaria relativizada.

Um melhor diagnóstico da qualidade da informação da cobertura censitária por idade pode ser feito analisando a informação anterior por sexo comparando-a com probabilidades de sobrevivência implícitas nos níveis de mortalidade da população estudada. O Gráfico 16.10 apresenta a RIS para cada sexo e se incluem probabilidades de sobrevivência correspondentes a níveis de mortalidade, fora dos quais, dificilmente localizar-se-ia a mortalidade brasileira[2].

[2] Para definir os intervalos foram usadas tábuas de vida modelo, concretamente o modelo "Oeste" de Princeton (ver Capítulo 20).

Gráfico 16.10: Brasil, 2000 e 2010: Razão Intercensitária de Sobrevivência por idade



Fonte: IBGE - SIDRA (https://sidra.ibge.gov.br/acervo#/S/Q) e Tabela A2 em Anexo.

A RIS por sexo para a população brasileira entre os anos 2000 e 2010 resulta numa curva que replica tanto o comportamento errático mencionado para a população total, como o sub-registro dos menores de 10 anos em 2000. Nos seguintes grupos de idade nota-se, para o sexo masculino uma diminuição da RIS das idades entre 15-19 e 25-29 relativamente mais acentuada que entre as mulheres. A hipótese adicional que surge é a de uma mortalidade masculina entre estas idades muito mais acentuada para os homens, fato que é bastante reconhecido. A seguir, se comparadas as curvas com as probabilidades de sobrevivência, que seguem um comportamento monotonamente decrescente (traços pontilhados), as RIS divergem. Tanto nos homens como entre mulheres aumentam ao ponto de se aproximar de 1,0 no caso dos homens e superar este valor no caso das mulheres.

Nas idades seguintes, a RIS masculina diminui muito; ao ponto de, nas idades 35-39 se identificar com níveis de mortalidade equivalente a $e_0 = 65$ anos. Note-se, também, que nas idades mais avançadas, o Censo indica RIS masculinas relativamente altas, se identificando com níveis de mortalidade com $eo = 80$ anos. Se por um lado, os níveis de mortalidade foram escolhidos arbitrariamente e implicam modelos cujo padrão pode não refletir necessariamente a situação brasileira, por outro lado, corroborariam a tendência das pessoas idosas a declarar ou ser declaradas com mais idade das que realmente teriam. No caso das mulheres, embora com menos intensidade, as variações da RIS com relação às probabilidades de sobrevivência consideradas, são similares às masculinas.

Em síntese, a RIS derivada destes dois censos, mostra um perfil que diverge de padrões esperados e que exige explicações; neste caso, os desvios sugerem:

- Diferenças de cobertura diferencial por idade entre os dois momentos censitários;
- Sub-registro do primeiro recenseamento principalmente nas primeiras idades;

- Sub-registro diferencial por sexo entre jovens no segundo momento, aumento da mortalidade juvenil masculina e/ou diferença abissal desta mortalidade por sexo em detrimento dos homens.

- Sub-registro das idades 20-25 anos no primeiro momento, ou (remotamente) imigração ao território brasileiro nestas idades.

16.5 A CONCILIAÇÃO CENSITÁRIA OU CONCILIAÇÃO DEMOGRÁFICA

Os métodos de avaliação e ajuste descritos na seção 16.3 funcionam bem quando os erros na informação são aleatórios ou pelo menos não implicam um aumento ou uma diminuição sistemáticos de certas características demográficas, como a idade das pessoas. Já as técnicas descritas na seção anterior e as que serão descritas abaixo lidam com erros que podem causar problemas mais sistemáticos na análise. Por exemplo, já foi mencionada a tendência sistemática à subdeclaração de crianças recém-nascidas nos censos. Outro problema é a tendência observada em muitos censos ao exagero das idades das pessoas mais velhas. Esse é um problema que não pode ser resolvido por meio de técnicas de ajuste de flutuações. As técnicas descritas na seção 16.4 podem ajudar a detectar problemas deste tipo. Por exemplo, uma comparação do grupo etário de 65-69 anos num censo com o grupo de 75-79 anos no próximo censo, dez anos depois, poderia revelar uma RIS improvavelmente elevada. A explicação poderia ser que boa parte das pessoas declaradas no segundo censo com idades entre 75 e 80 anos em realidade poderiam ter 70-74 anos. Este erro seria parcialmente compensado pela transferência de outras pessoas realmente pertencentes à faixa de 75-79 para a faixa seguinte, de 80-84. Entretanto como a população nesta fase da vida diminui rapidamente com a idade, essa compensação seria apenas parcial e haveria um exagero líquido do tamanho da faixa de 75-79 anos, resultando num RIS improvável entre as faixas de 65-69 e 75-79. A correção deste tipo de distorções é mais difícil do que o ajuste das flutuações aleatórias tratado na seção 16.3. Além das técnicas descritas abaixo, às vezes pode ser útil recorrer a modelos formais que descrevem a estrutura etária de uma população e que serão discutidos no Capítulo 22.

A técnica mas frequentemente usada para a correção de erros sistemáticos é a denominada conciliação demográfica cujo objetivo básico, tal como explicita IBGE (2008) é "*aferir os níveis esperados de coerência entre a informação dos censos e os eventos demográficos – nascimentos, mortes e migração – de tal maneira que se cumpra (ou que se aproxime ao máximo) o explicitado na conhecida equação compensadora*". A utilidade principal da conciliação censitária se manifesta na estimação de uma população base para a preparação de projeções de população. Essa população base precisa refletir não só o último dado censitário disponível, mas precisa ser consistente com toda a história demográfica conhecida antes do último censo.

É claro que muitas ambiguidades surgem com relação à confiabilidade das fontes a serem usadas em cada contexto e à exatidão dos resultados obtidos. Em todo caso, uma conciliação censitária, isto é, a obtenção de informação por sexo e idade de uma dada coorte que seja coerente entre os diversos censos, incluiria, como definido por Rincón (1984 b), quem faz uma didática e detalhada descrição desta técnica, as seguintes fases:

- Avaliação do grau de cobertura de cada um dos censos demográficos. Esta avaliação é feita, inicialmente, utilizando os resultados de pesquisas de avaliação da cobertura feitas imediatamente após a realização das entrevistas (ver seção 4.4 do Capítulo 4).

- Correção das distribuições por sexo e idade dos censos no que toca à falta de cobertura, subenumerações diferenciais e má declaração da idade. Isto pode ser feito mediante reconstituição das diversas coortes implícitas nos grupos etários utilizando, por exemplo informação sobre o número de nascimentos, óbitos e movimentos migratórios. Esta informação pode ser obtida diretamente de registros vitais se existem e são confiáveis, e/ou mediante o uso de técnicas de estimação indireta. No caso dos nascimentos, a ausência de registros vitais confiáveis pode ser substituída por estimativas de fecundidade derivadas do mesmo censo; o mesmo se aplica às estimativas de mortalidade e particularmente, no caso de movimentos migratórios. Como se sabe, não é comum a existência de registros contínuos de deslocamentos populacionais, devendo, neste caso, acudir a técnicas de medição destes movimentos a partir dos dados censitários.

- Compatibilização da dinâmica demográfica de dois ou mais períodos intercensitários buscando verificar a coerência dos censos com as estimativas da mortalidade, da fecundidade e da migração, considerando o máximo de informações disponíveis e confiáveis, aqui incluídos, o uso de técnicas indiretas de análise para a fecundidade a mortalidade e a migração e estatísticas contínuas de diversos grau de confiabilidade para obter diretamente as estimativas necessárias.

A confiabilidade dos resultados após feita a conciliação censitária dependerá da intensidade do grau de cobertura/subenumeração estimado e da coerência entre os censos adjacentes. O diagnóstico final terá sempre algum grau de subjetividade analítica.

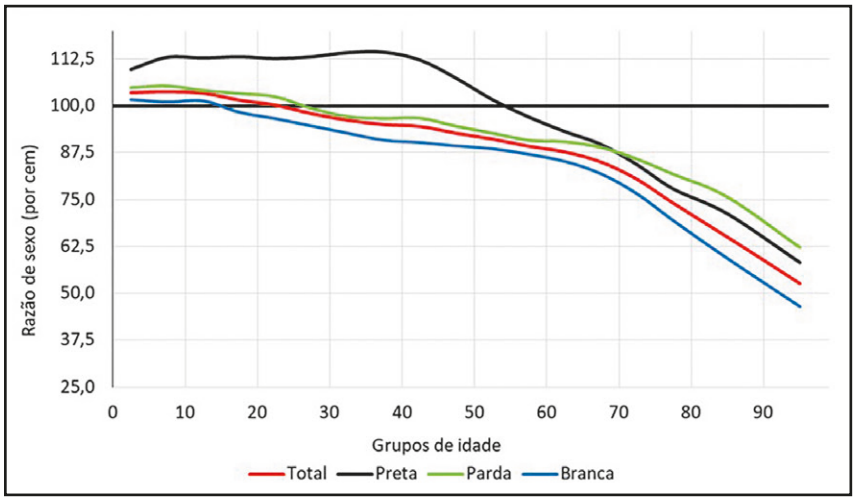
16.6 AS VARIÁVEIS SEXO E IDADE E A FECUNDIDADE, MORTALIDADE E MIGRAÇÃO

Erros na declaração do sexo são mais raros do que erros na declaração da idade. Como se viu no Capítulo 6, a razão de sexo ao nascer (RS_0) geralmente se situa entre limites relativamente estreitos e quando sai do intervalo normal, as razões mais prováveis são uma distorção real do número de nascimentos femininos devido ao aborto seletivo ou a subdeclaração desse número. Entretanto, também há exemplos de erros mais complexos, como o seguinte, derivado da informação sobre população por sexo, idade e cor no Censo brasileiro de 2010.

O Gráfico 16.11 apresenta as *RS* desagregadas por idade e cor. O comportamento da população total reproduz o padrão por idade já descrito anteriormente e se aplica à população *branca* e, também, *parda* embora a RS_x para estas duas categorias fique constantemente acima e abaixo da média respectivamente sendo que no caso da população branca, a RS_{0-4}, o valor fica muito próximo a 100 (linha negritada, no gráfico). Chama a atenção o comportamento da RS_x da população *preta*. A RS_{0-4} é de 110 e se mantém alta até a idade 35-39 quando a *RS* atinge seu valor máximo de 114 para declinar logo muito acentuadamente, ao ponto de, a partir da idade 40, ser menor que a da população parda. Esta diferenciação por cor em RS_{0-4} é semelhante à encontrada no registro de nascimentos.

Os determinantes deste incomum comportamento, já encontrado no Censo anterior, não são totalmente compreendidos, mas parece improvável que o padrão se deva a uma seletividade por sexo nos nascimentos na população negra, mesmo porque o desequilíbrio inicialmente aumenta com a idade. Uma explicação mais provável é que o padrão se deva a uma maior tendência a declarar bebês e crianças de sexo feminino não brancas como brancas ou pardas, antes que pretas.

Gráfico 16.11: Brasil, 2010 – Razões de sexo por idade segundo cor da pele



Fonte: IBGE - SIDRA (https://sidra.ibge.gov.br/acervo#/S/Q).

Abaixo se discutem brevemente os erros possíveis nas componentes demográficas. Para uma discussão mais completa do assunto, pode-se consultar a Terceira Parte do manual publicado pelo CELADE (Naciones Unidas, 2014). Quando se trata de estudar os aspectos da dinâmica demográfica, a medição da fecundidade, mortalidade e migração deve considerar que os padrões de erro na declaração do sexo e idade e os correspondentes ajustes tanto no numerador (eventos) como no denominador (população) podem ser diferentes pois a origem das respostas e a forma de correção/ajuste são também diferentes.

Na medida em que os vieses por sexo e idade no denominador sejam diferentes dos do numerador, diferentes serão os erros nas medidas de fecundidade, mortalidade ou migração. A seguir, alguns exemplos sobre idade e sexo e vieses na:

Mortalidade

- Nos casos em que a cobertura do registro de óbitos é incompleta, é conhecida a tendência de maior omissão de óbitos menores de um ano, sendo, geralmente, maior a omissão quanto menor é a idade da criança, principalmente se o óbito ocorre nos primeiros instantes de vida, quando o nascido vivo (nado vivo) pode ser erroneamente registrado como nascido morto (nado morto). A avaliação da omissão de óbitos de crianças pode ser feita mediante a aplicação de métodos indiretos de análise demográfica, como o método dos filhos sobreviventes de Brass (1975). Em se tratando de estimar taxas de mortalidade por idade, usando a informação censitária (denominador) e a informação de estatísticas vitais (numerador), é

comum encontrar maior omissão nesta última. Sem as necessárias correções o resultado é a subestimação das taxas de mortalidade. O ajuste da informação sobre a população pode ser feito, por exemplo por meio do processo de conciliação demográfica e o ajuste no caso dos óbitos pode ser feito, por exemplo, mediante a aplicação de técnicas indiretas (ver Capítulo 23) ou diretas, como os processos de busca ativa de eventos (Szwarcwald et al., 2011).

- No caso de omissão da declaração de meninas nas primeiras idades, deve-se estar alerta sobre as consequências no diferencial por sexo da mortalidade na infância que isto causa. O registro das mortes, por ser compulsório em determinados contextos, causaria sobreestimação da taxa de mortalidade feminina nestas idades em razão da subestimação do denominador, influenciando, assim, o esperado comportamento diferencial da mortalidade por sexo.

- No caso de declaração tendenciosa por sexo e idade de características selecionadas, como é o caso de cor/raça, no Brasil, estimativas de mortalidade por tais características, terão também, sérios vieses dependendo dos mecanismos de definição desta característica e dos preconceitos sociais prevalecentes ao respeito – seja no numerador ou denominador. Lembrar que no caso do denominador (população), a informação é fornecida pelo indivíduo entrevistado se tratando, frequentemente de autoeclaração. No caso do numerador (óbito) o informante será sempre uma terceira pessoa (médico, parente, autoridade forense, dependendo da situação).

Fecundidade

- O cálculo de taxas de fecundidade por idade a partir de informação censitária relacionando as mulheres com os prováveis filhos (como no caso do método dos filhos próprios, ver Capítulo 23) é um caso claro de ocorrência de vieses diferentes na declaração da idade no numerador e no denominador e a possibilidade de produzir estimativas erradas de fecundidade. Sobre o denominador: a omissão de mulheres pode se dar mais acentuadamente nos extremos do período reprodutivo, como se viu, e, além do mais, existe a tendência ao arredondamento ou preferência por certos dígitos. Sobre o numerador: entre as crianças (menores de 5 ou 10 anos) a omissão tende a ser mais acentuada entre as mais jovens (menores de um anos e menores de um mês) principalmente se o nascido vivo (nado vivo) faleceu; e menos acentuada ao aproximar-se da idade 10. O padrão de erro na declaração de idade e a atração por dígitos ou arredondamento são muito menos acentuados dada a grande diferença física entre crianças com idade, digamos, 5 e 10 anos.

- Informação sobre sexo dos filhos tidos no(s) último(s) ano(s) prévio(s) à data do recenseamento: Sabendo que há uma razão de sexo constante ao nascer, esta deve se manter segundo a idade da mãe e qualquer outra característica a ser considerada (educação, renda, divisão administrativa etc.). Esta informação serve como parâmetro da qualidade da informação e guia para avaliar a necessidade de eventuais ajustes. Deve lembrar-se que poderão existir situações nas quais, uma razão de sexos fora do normalmente esperado pode ser real; este seria o caso de seletividade por sexo de abortos onde há forte preferência por crianças de determinado sexo.

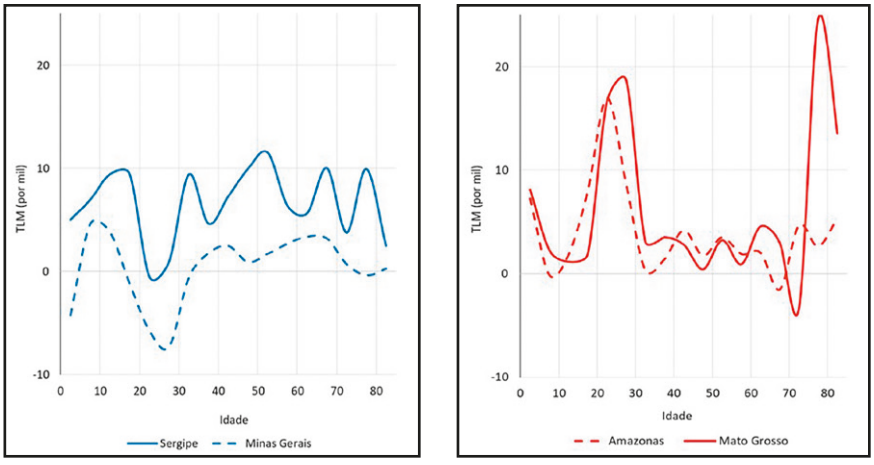
- Parturição por idade: O número de filhos tidos nascidos vivos (ou parturição) é uma variável de estoque e por esta razão, quando obtida num censo, refere-se a informação retrospectiva, cuja confiabilidade costuma ser afetada por erros de memória que se acentuam quando aumenta a idade da mulher. Este viés costuma, ademais, estar associado ao status socioeconômico da mulher. Em contextos rurais ou onde a mulher tem muito pouco acesso à educação, a omissão por idade, costuma se acentuar. Esta é uma das razões pela qual o método P/F de Brass não usa as parturições das mulheres no extremo das idades reprodutivas para aferir níveis de fecundidade.

Migração

- Na aferição dos movimentos migratórios é preciso ter em conta o impacto dos padrões de erro na declaração por idade e na temporalidade dos fluxos, pois na basta maioria dos casos, a informação é retrospectiva e deriva-se dos censos. O cálculo de fluxos migratórios por idade e tempo de residência ilustra bem esta situação: A declaração da idade como se viu, pode estar viesada pela atração de dígitos ou arredondamentos. A resposta sobre o número de anos residindo no lugar atual de residência, além de estar sujeita a arredondamentos e atração de dígitos está atrelada à data do recenseamento. Exemplo: o informante pode declarar que tem 30 anos (arredondamento) e declarar que mora no lugar de residência 5 anos (arredondamento). A estimativa da data do movimento migratório será provavelmente em torno desses 5 anos antes do censo; este viés conjuntamente com aquele sobre a idade com que foi feito o movimento impactará as estimativas obtidas.

- Os padrões diferenciados de erro na declaração da idade em dois censos consecutivos, influenciam também, as medidas de migração. O Gráfico 16.12 mostra uma estimativa da Taxa Líquida de Migração por idade de quatro UFs brasileiras de diversos graus de desenvolvimento socioeconômico. No lado direito apresentam-se estimativas para a população feminina, no lado esquerdo, para a população masculina.

Gráfico 16.12: Brasil, 2005-2010: Taxas líquidas migratórias para homens e mulheres, por idade para quatro Unidades Federativas selecionadas: Sergipe e Minas Gerais (homens) e Amazonas e Mato Grosso (mulheres)



Fonte: CEDEPLAR, Laboratório de Estimativas Demográficas (2015).

É pouco provável que o padrão por idade da migração que antecedeu ao quinquênio censitário correspondente siga o zigue-zague apresentado no gráfico, seja para homens ou mulheres. Como as estimativas foram obtidas a partir da informação sobre residência e idade, comparada nos dois censos consecutivos, uma das explicações para grande parte do comportamento errático das curvas é a falta de coerência na distribuição por idade da população entre os censos.

É necessário, consequentemente, se a distribuição por idade da população registrada nos censos não foi ajustada, ajustar a distribuição por idade das taxas líquidas migratórias. O processo inclui de técnicas de modelagem procurando suavizar o padrão atendendo, claro, ao contexto no qual se insere o processo migratório (Rogers e Castro, 1981; ver também o Capítulo 20).