INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO
UNIVERSIDADE TÉCNICA DE LISBOA

FALTA DE RESPOSTA EM SONDAGENS

Dissertação apresentada como requisito parcial para a obtenção do grau de mestre em Matemática Aplicada à Economia e à Gestão

Ana Isabel Gonçalves da Costa Lorga da Silva

Junho de 1995

Agradecimentos

Em primeiro lugar, agradeço ao Professor Doutor Bento Murteira, a sua valiosa orientação e disponibilidade, que foram fundamentais para a realização deste trabalho, sendo, no entanto ,evidente que quaisquer erros ou omissões são da minha total responsabilidade.

Ao Dr. Arnaldo Lopes, ao Dr. Fernando Casimiro e à Dra. Helena Gomes, do I.N.E., a colaboração na disponibilização de informação prática.

Aos meus colegas das equipas de Matemática I e de Matemática II, do ano lectivo de 1992/93, a colaboração que me deram ao longo desse ano em que realizei a parte escolar do mestrado, assim como às minhas colegas de mestrado Alécia Vale e Isabel Reis.

Aos meus colegas de equipa de Matemática I do ano lectivo de 1993/94, o incentivo que me deram no início da realização deste trabalho.

Aos meus pais e marido a sua presença, compreensão e carinho muito especiais, nos momentos necessários.

Aos meus colegas de gabinete, o seu apoio.

Finalmente, a cada um dos meus amigos, a contribuição individual.

INDICE

CAPÍTULO 1

INTRODUÇÃO

A análise estatística, quando aplicada a um conjunto de n unidades estatísticas e de p variáveis, $Y_1, Y_2, \ldots, Y_p$, considera como base uma matriz de dados $n \times p$,

(1.1) $$\left[Y_{ij}\right], i = 1,2,\ldots,n; j = 1,2,\ldots,p,$$

em cujas linhas se situam as n unidades inquiridas (objectos, pessoas, famílias,. empresas, cobaias, etc.) e em cujas colunas se situam as observações sobre cada unidade das p variáveis (contínuas, ordinais e nominais) consideradas.

Em qualquer caso existe normalmente a preocupação de ter presente a amplitude do processo que conduziu aos dados. Trata-se de uma **sondagem** ou inquérito por amostragem ou trata-se de um **censo** ou indagação exaustiva. Tratando-se de uma sondagem interessa evidentemente também registar qual o tipo de amostragem adoptado para recolher os dados (amostragem simples, estratificada, por conglomerados, etc.).

A partir da matriz de dados, em função dos objectivos do estudo e com recurso a *packages* estatísticos mais ou menos poderosos (SAS, SPSS, etc.), aplicam-se os mais variados procedimentos (cálculo de médias, variâncias, regressão múltipla, correlação canónica, componentes principais, etc.) e procede-se à análise das conclusões a que conduzem.

Convém, porém, não esquecer que a maioria dos procedimentos estatísticos, nomeadamente os referidos acima, trabalham quase sempre sobre uma **matriz completa.** Põe-se então o problema de saber como se chegou a uma matriz sem lacunas.

Deu-se a feliz e rara circunstância de terem "respondido" de forma cabal todas as unidades designadas pelo esquema de amostragem para fazer parte da amostra ou todas as unidades designadas (por norma ou por lei) para ser objecto de recenseamento?

Em alternativa:

Foram riscadas ou ignoradas as unidades que não responderam de todo ou responderam de forma parcial?

Ou foram de algum modo atribuídos valores às variáveis das unidades em falta ou aos itens em falta respeitantes a algumas das variáveis?

No domínio da análise estatística, a verdade é que ao chegar ao fim da fase de recolha e de codificação dos dados, as matrizes completas são a

excepção e não a regra. A falta de resposta ou a existência de dados perdidos ("*missing data*") é, pode dizer-se, uma quase fatalidade quando se encara a fria realidade dos inquéritos por amostragem ou dos censos, sobretudo nos estudos sociais em que se inquirem pessoas, famílias, empresas ou instituições. Como dizem Little e Rubin(1987), num inquérito em que as unidades são famílias, é natural que algumas se recusem a declarar o rendimento; numa experiência industrial pode haver dados perdidos por haver falhas mecânicas nos processos em estudo; numa sondagem à opinião pública os indivíduos inquiridos podem recusar-se a responder ou não ter qualquer preferência.

Existe, evidentemente, um grande leque de **medidas preventivas** que procuram evitar tanto quanto possível a existência de lacunas na matriz dos dados, isto é, que procuram reduzir o número de não respondentes ou de dados perdidos[1]. No entanto, esgotadas as potencialidades de tais medidas, verifica-se quase sempre a necessidade de completar a matriz, isto é, passar da matriz inicial, em que faltam unidades ou faltam itens, para a matriz completa. Para o efeito pode seguir-se a abordagem que consiste em eliminar todas as unidades que não responderam ou responderam de forma incompleta e passar adiante estimando médias, totais, etc, etc, ou calculando componentes principais, correlações canónicas, etc, etc. É a

[1] Lessler e Kalsbeek(1992) considerando diferentes modalidades de inquérito (por correio, por telefone, por entrevista, etc) referem, entre outros, os seguintes métodos preventivos: *(a) high-priority mail, (b) more call attempts, (c) clearer interview assignment materials, (d) special tracing efforts, (e) follow-up reminders, (f) incentives, (g) endorsements, (h) use of female interviewers, (i) lead letter, (j) proxy respondents e (k) refusal conversion strategies.*

abordagem "*do nothing*" muitas vezes seguida, por ignorância ou por se ter consciência de que a taxa de resposta é elevada ou de que há poucos casos de dados perdidos. Contudo, como vai mostrar-se, pode dar lugar a enviesamentos ou ineficiências nomeadamente quando as unidades não respondentes, ou com dados perdidos, representam uma proporção significativa e têm características diferentes das unidades respondentes.

Em termos muito gerais, que mais tarde se concretizam um pouco, existem formas mais sofisticadas de completar a matriz. Essas formas podem eventualmente variar consoante se trate de uma sondagem ou de um censo. Tendo em vista a superficialidade das considerações introdutórias não se faz, por enquanto, qualquer distinção.

Assim, pode recorrer-se à **extrapolação** através da qual se projecta para os não respondentes estimativas obtidas observando o que se passa com os respondentes ordenados em função do respectivo grau de semelhança com os não respondentes.

Pode ainda recorrer-se a uma **subamostra de não respondentes**, desde que esta tenha garantia de sucesso, tirando conclusões ou fazendo estimativas a partir do subconjunto dos respondentes e da subamostra de não respondentes. Trata-se de um procedimento que envolve amostragem dupla.

Pode também recorrer-se a métodos de **imputação** em que os valores em falta são prenchidos através de grande número de critérios, entre os quais se destacam a título de exemplo:

(1) **imputação da média** em que a média dos valores registados no grupo dos respondentes (ou num subgrupo de características semelhantes às do não respondente) toma lugar do item em falta;

(2) **imputação explícita,** por exemplo, **imputação por regressão** em que a partir de variáveis auxiliares que são conhecidas para os não respondentes se faz a predição dos valores em falta por meio de um modelo de regressão;

(3) **imputação "*hot deck*"** em que com certa lógica os valores observados tornam-se "dadores" em relação aos casos em falta.

Pode igualmente recorrer-se a métodos de **ponderação** em que na construção de estimadores os valores observados são ajustados ou ponderados de modo a conseguir compensar a falta de resposta.

Pode enfim recorrer-se a **procedimentos baseados em modelos.** Trata-se de uma vasta classe de soluções obtidas através da definição de um modelo cujos parâmetros podem estimar-se empregando os dados dos

valores em falta ou eventualmente dos valores do universo (finito) correspondentes a unidades não incluídas na amostra. Em algumas das soluções tem papel importante a inferência Bayesiana.

Já foi dito que a falta de resposta ou a existência de dados perdidos é uma quase fatalidade quando se encara a fria realidade dos inquéritos por amostragem ou dos censos. E é uma quase fatalidade que pode ter graves consequências em relação à validade das conclusões das sondagens ou dos censos. Apesar da gravidade do problema pode dizer-se que só nos anos 70 começou a aparecer literatura aprofundando os problemas levantados pela análise estatística de conjuntos de observações com dados em falta ou dados perdidos.

Na primeira edição do seu clássico livro - "*Sampling Methods for Censuses and Surveys*" - Yates(1949) não deixa de fazer uma advertência quanto aos efeitos da não resposta: "*Unless non-response is confined to a small proportion of the whole sample the results cannot claim any general validity*", mas não vai muito além nos remédios propostos (*follow-up*, subamostra dos não respondentes e pouco mais), nem deixa de afirmar que: "*The simplest way of dealing with non-response is to regard the*

non-respondents as similar to the remainder of the sample, i.e., to treat the sample as if it were a sample on a smaller number of units.".

Era assim em meados do século, e Yates, estatístico notável que era, reflecte necessariamente o "estado da arte". De Yates para o não menos clássico Hansen, Hurwitz e Madow(1953) há algum progresso sobretudo no domínio da amostragem dupla. Mas, como dizem Little e Rubin(1987), foi o avanço no domínio da computação que permitiu, a partir de 1970, o florescimento de uma abundante literatura sobre a não resposta e os dados perdidos que culminou em 1983 com a publicação[1] da importante obra em três volumes: "*Incomplete Data in Sample Surveys*":

Vol.1 - *Reports and Case Studies* - Ed. William G. Madow, Harold Nisselson e Ingram Olkin;

Vol.2 - *Theory and Bibliographies* - Ed. William G. Madow, Ingram Olkin e Donald B. Rubin;

Vol.3 - *Proceedings of the Symposium* - Ed. William G. Madow e Ingram Olkin.

[1] Academic Press, Nova Iorque.

Não é objectivo do presente trabalho, nem tal seria possível, apreciar o enorme fluxo de contribuições contido nessa obra e nos desenvolvimentos posteriores. Dada a complexidade da questão, o objectivo que se pretende é bem mais elementar. Entende-se justificada mais uma chamada de atenção para as importantes consequências da não resposta em sondagens e censos e para alguns dos métodos propostos para lidar com o problema. Pretende-se também dar uma modesta contribuição para combater a excessiva tendência para aplicação da técnica da "avestruz" ou "*do nothing*", isto é, a prática de meter a cabeça na areia e proceder como se os dados estivessem completos depois de os expurgar das unidades não respondentes ou com respostas incompletas. Complementarmente pretendem apresentar-se algumas soluções para o problema praticadas no País, mais concretamente no I.N.E..

CAPÍTULO 2

ERROS EM SONDAGENS

Depois de por volta de 1895 Krier, numa reunião do *International Statistical Institute*, ter declarado, com alguma celeuma, que preferia "*a small number of careful observations carried out with great care to a large number of superficial observations made superficially on a large scale*", tem-se verificado um grande aumento na realização de sondagens e de inquéritos por amostragem nos mais diversos contextos. Entre nós, o incremento, na parte a que o grande público tem acesso, está bem patente quer, por um lado, na profusão de inquéritos regularmente publicados pelo I.N.E. (Inquérito Mensal à Construção Civil e Obras Públicas, Inquérito Mensal à Indústria Transformadora, Inquérito de Conjuntura aos Serviços, Inquérito de Conjuntura ao Investimento, entre muitos outros), quer, por outro lado, no sistemático recurso a sondagens por parte dos meios de comunicação social.

A tendência que se observa, se representa uma confiança mais generalizada nas inferências feitas com base em amostras recolhidas de um conjunto mais vasto - o universo, resulta também do desenvolvimento dos

métodos estatísticos, nomeadamente os associados com a chamada teoria da amostragem, e do progresso registado nos meios de computação, com particular ênfase para o vasto *software* disponível para aplicação da mesma teoria.

A par deste desenvolvimento existe uma crescente tomada de consciência de que podem encontrar-se erros nas sondagens, facto que tem levado à criação de métodos com os quais se procura de forma eficiente efectuar o controlo de tais erros.

De um modo geral é difícil definir o que se entende por **erros em sondagens**. Pode afirmar-se, de forma algo vaga, que semelhantes erros ocorrem quando existe discrepância entre as conclusões e a realidade; no entanto, a existência desta discrepância e a sua avaliação pode ser altamente controversa. O mais aconselhável consiste, na verdade, em proceder a uma esquemática inventariação dos diferentes tipos de erro possíveis os quais naturalmente se agregam para dar uma resultante.

2.1 - Erros de amostragem e erros não provocados por amostragem

O título da presente secção introduz já uma dicotomia particularmente elucidativa. Dizem-se **erros de amostragem** os que resultam da decisão consciente de estudar uma amostra ou subconjunto, de dimensão n, em vez de estudar exaustivamente a população ou universo - todo - de dimensão

$N > n$. Trata-se de um tipo de erro inerente ao próprio processo de amostragem e, portanto, inevitável. Quando a amostra é recolhida de acordo com os métodos probabilísticos consagrados os erros de amostragem podem controlar-se. Quer dizer, com os métodos probabilísticos consegue-se assegurar, tendencialmente[1], a representatividade da amostra na medida em que se evitam os enviesamentos correntemente introduzidos quando a recolha se baseia inteiramente no critério de escolha do observador.

Logicamente, os **erros não provocados por amostragem** são aqueles que não se encontram associados com a inevitável incerteza decorrente do processo que consiste em realizar inferências sobre a população a partir da amostra. Mais concretamente são erros devidos inteiramente a deficiências ocorridas durante a execução dos procedimentos que compõem o plano de amostragem. Podem resultar dos mais diversos factores tais como definições erradamente concebidas, imperfeições na esquematização do próprio plano de amostragem e na base de amostragem ou *frame*[2] utilizado, insucesso na recolha de dados ou respostas, erros de medida ou de

[1] Nenhum método probabilístico pode assegurar a representatividade de **uma** particular amostra. Por exemplo, lançando um dado ao acaso 10 vezes não é impossível (mas altamente improvável) obter 10 vezes a face marcada com 6 pontos. A teoria clássica raciocina em termos de amostragem repetida e é um facto que a grande maioria das amostras casuais é, em geral, representativa.

[2] Por exemplo, no Inquérito ao Transporte Rodoviário de Mercadoria (I.N.E., 1990), no que se refere ao Continente, a base de amostragem ou *frame* foram os ficheiros da Direcção-Geral de Transportes Terrestres, constituídos a partir das atribuições dos impostos de circulação e camionagem, respectivamente para o parque particular e parque público. Pode dizer-se que o *frame* é o mecanismo (em muitos casos uma lista ou ficheiro) que permite o acesso à população.

computação ao trabalhar os dados recolhidos, etc., etc. Esquematicamente,

$\Rightarrow$ erros na base de amostragem ;

$\Rightarrow$ erros de não resposta;

$\Rightarrow$ erros de medida ou de cálculo,

são os principais erros não provocados por amostragem.

Uma vez que os erros não provocados por amostragem podem explicar a maior parte do erro total da sondagem, mesmo atendendo a que a sua grandeza não pode muitas vezes ser medida, é importante que quem elabora o plano da sondagem procure assegurar a implementação de processos para os controlar e reduzir na medida do possível.

Assim, quando um aspecto importante é a estimação de alguma característica da população, por exemplo, a média do rendimento familiar, seja $\overline{Y}$, através da média da amostra, seja $\overline{y}$, o *design* ou planeamento da sondagem deve traduzir o esforço feito para controlar o erro total de estimação qualquer que seja a sua origem. O erro quadrático médio (*EQM*) é a expressão mais simples e mais frequentemente ulizada para medir esse

erro,

$$EQM(\bar{y}) = \sum_k \frac{S_k^2}{m_k} + (\sum_k B_k)^2. \tag{2.1}$$

Na expressão acima, a primeira parcela representa a variância dos erros de todas as k fontes e a segunda representa o quadrado da soma dos enviesamentos associados com as k fontes; m_k designa o número de unidades.

Um aspecto interessante da fórmula (2.1) - de Kish(1965) - é mostrar que enquanto as variâncias podem reduzir-se aumentando a dimensão da amostra tal aumento deixa os enviesamentos insensíveis.

2.2 - Fontes de erros não provocados por amostragem

Apesar da esquemática classificação dos erros não provocados por amostragem julga-se conveniente alargar um pouco as considerações sobre as fontes de tais erros.

De um modo geral, é muito difícil, ou até mesmo impossível, fazer uma listagem das fontes de erro em sondagens, isto é, de todos os factores que de algum modo possam contribuir para a introdução de erros.

Do que foi exposto anteriormente, pode já afirmar-se que podem contribuir para a presença de erros uma má definição ou inexequibilidade dos objectivos da sondagem e uma deficiente caracterização da população que pretende sondar-se. Outro factor importante e eventualmente limitativo envolve o aspecto monetário, quer dizer, a quantia disponível para executar a sondagem. Factores também relevantes são o modo de seleccionar a amostra, de recolher e processar os dados. Para a maior ou menor gravidade de tais erros contam enormemente os conhecimentos e a experiência dos responsáveis pela realização da sondagem.

O efeito de uma fonte de erro depende do potencial próprio para causar o erro e do grau com que pode ser controlada. Uma fonte pode ter grande potencial mas provocar poucos erros se for eficientemente controlada.

Quando pretende fazer-se uma sondagem, por que se entende ser o melhor método para obter a informação desejada, nunca é demais insistir na importância de um cuidadoso planeamento com particular destaque para as questões a levantar, para a selecção de meios humanos qualificados e para a orçamentação dos custos.

Voltando ao aspecto monetário, pode dizer-se que um orçamento insuficiente pode comprometer seriamente o sucesso da sondagem. Por exemplo, quando a recolha é feita por entrevista pessoal, diminuir por

questões orçamentais o número de entrevistadores pode ter como consequência uma redução na qualidade dos dados. O entrevistador pode não ser tão persistente, por falta de tempo, ao solicitar a colaboração dos entrevistados que podem assim deixar de responder a questões que noutras circunstâncias talvez fossem respondidas.

Quando os inquiridores não compreendem bem as normas definidas pela organização, os resultados da sondagem podem ser gravemente comprometidos.

Pode também ser prejudicial para os resultados o facto de alguns entrevistadores serem mais convincentes do que outros ao solicitarem a colaboração dos entrevistados, facto que provoca desníveis em termos de respostas.

A boa compreensão dos entrevistados quanto à natureza do inquérito e o bom entendimento das questões postas é, também, um factor relevante na minoração de erros. Designadamente deve ter-se em atenção a linguagem utilizada, pois quando a sondagem envolve termos técnicos - por exemplo, acerca da nova tecnologia automóvel - o questionário pode não ser compreendido por todos os indivíduos e, assim, uns por vergonha, outros apenas por vontade de colaborar, outros por motivos distintos, podem responder mesmo sem perceberem muito bem o que estão a dizer ou eventualmente recusar-se a responder.

Embora a maioria das sondagens utilize métodos probabilísticos para seleccionar as unidades a incluir na amostra, podem surgir erros do tipo que consiste em seleccionar moradas distintas que correspondem à mesma família, o mesmo podendo suceder com números de telefone.

Outros problemas habituais decorrem da dificuldade, ou mesmo da impossibilidade, de localizar as unidades a entrevistar e da estrutura do questionário.

Quando se projecta um questionário deve ter-se em atenção a ordem e o modo como se colocam as questões para evitar sugestionar os inquiridos na forma como respondem. Por exemplo, quando se indaga acerca das preferências por marcas de gelados, se na entrevista pessoal ou telefónica não se obtem uma resposta imediata, a enumeração de marcas por parte do entrevistador pode influenciar o entrevistado. Analogamente, se o entrevistador coloca as questões dizendo "Não pensa que ...", o entrevistado pode ser arrastado a dar uma resposta diferente da que daria se a questão tivesse sido posta na forma "Diga o que pensa sobre ...".

Um dos factores que contribui para que os inquiridos não colaborem conscientemente no preenchimento completo de um questionário é a respectiva extensão; quando o questionário é muito extenso os inquiridos podem desistir de responder a partir de certa altura ou dar mesmo resposta mais ou menos ao acaso.

Finalmente, não podem ignorar-se os problemas que podem surgir ao nível da codificação das respostas e no tratamento estatístico e informático dos dados. As actuais potencialidades de computação podem não evitar erros contidos na base de dados ou a aplicação de métodos de tratamento que não sejam os mais adequados.

CAPÍTULO 3

NÃO RESPOSTA EM SONDAGENS

Da longa e complexa descrição da natureza e das fontes de erros não provocados por amostragem facilmente se conclui que a falta de resposta é um dos mais importantes e é, aliás, aquele de que se ocupa o presente trabalho.

De um modo genérico designa-se por **não resposta** ou **falta de resposta** a não existência de dados sobre uma parte da amostra. No caso da matriz de dados referida no capítulo 1 essa inexistência caracteriza-se por uma matriz incompleta. A não resposta ou falta de resposta são erros que resultam claramente de uma situação de insucesso na recolha de dados.

3.1 - Distinção entre unidade e item de não resposta

Quando se fala em não resposta devem distinguir-se duas situações:

* **unidade de não resposta** - se um elemento da amostra é abordado e não se obtem qualquer resposta, isto é, se há uma unidade amostral da qual se afirma que há falta de resposta;

* **item de não resposta** - se um elemento da amostra é solicitado a responder, colabora, mas não fornece toda a informação necessária, caso em que há pelos menos um item com falta de resposta.

Exemplificando: em sondagem que consista em aplicar um questionário a determinado número de empresas, se uma empresa se recusa a responder ao questionário diz-se que é uma unidade de não resposta; se uma empresa se não recusa a responder, mas não responde a uma (ou mais do que uma) questão está-se perante um item de não resposta.

Suponha-se que de uma população de $N=100$ empresas (unidades) é retirada uma amostra de dimensão $n=9$ e que a sondagem emprega um questionário composto por cinco questões (itens). Pretende-se que as

nove unidades amostrais respondam a todos os itens para obter a matriz completa:

$$\begin{bmatrix} Y_{11} & Y_{12} & Y_{13} & Y_{14} & Y_{15} \\ Y_{21} & Y_{22} & Y_{23} & Y_{24} & Y_{25} \\ Y_{31} & Y_{32} & Y_{33} & Y_{34} & Y_{35} \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} & Y_{45} \\ Y_{51} & Y_{52} & Y_{53} & Y_{54} & Y_{55} \\ Y_{61} & Y_{62} & Y_{63} & Y_{64} & Y_{65} \\ Y_{71} & Y_{72} & Y_{73} & Y_{74} & Y_{75} \\ Y_{81} & Y_{82} & Y_{83} & Y_{84} & Y_{85} \\ Y_{91} & Y_{92} & Y_{93} & Y_{94} & Y_{95} \end{bmatrix}$$

Mas se se obtem resposta apenas de oito empresas - oito unidades respondentes - e se por exemplo há duas que não respondem a uma questão e há uma que não responde a duas questões - têm-se quatro itens de não resposta:

$$\begin{bmatrix} Y_{11} & Y_{12} & Y_{13} & Y_{14} & Y_{15} \\ Y_{21} & Y_{22} & Y_{23} & Y_{24} & Y_{25} \\ Y_{31} & - & Y_{33} & Y_{34} & - \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} & Y_{45} \\ - & Y_{52} & Y_{53} & Y_{54} & Y_{55} \\ Y_{61} & Y_{62} & Y_{63} & Y_{64} & Y_{65} \\ Y_{71} & Y_{72} & Y_{73} & Y_{74} & - \\ Y_{81} & Y_{82} & Y_{83} & Y_{84} & Y_{85} \\ - & - & - & - & - \end{bmatrix}$$

3.2 - Motivos para a não resposta

A taxa de resposta está associada à característica estudada.

Os erros de não resposta podem surgir por diversas razões, tais como a natureza da informação observada, as características dos indivíduos e o método utilizado para recolher informação. Estes factores estão com frequência inter-relacionados.

Lessler e Kalsbeek(1992) classificam os motivos para não resposta em cinco categorias. As duas primeiras são:

Não eleito ("*Ineligible*") - unidades que foram seleccionadas, mas que não são incluídas, por estarem fora do âmbito da sondagem.

Não solicitado ("*Not Solicited*") - em que as unidades não são localizadas, e como consequência a solicitação para obtenção de resposta não é conseguida.

A terceira - **solicitado mas indisponível** ("*Solicited but Unwilling)* - refere-se ao caso em que as unidades não respondentes são localizadas e solicitadas, mas não colaboram, ou ao caso de item de não resposta, em que não respondem a determinadas questões. Enquanto na terceira categoria a falta de resposta é atribuída aos indivíduos que decidem em

consciência não responder, por medo, apatia, desconfiança ou falta de tempo, na quarta - **solicitado mas impossibilitado** ("*Solicited but Unable*") - a falta de resposta é atribuída à incapacidade de resposta por motivos físicos, mentais, emocionais ou problemas de linguagem.

Finalmente, os restantes motivos - **outros motivos** ("*Other Reasons*") - incluem marginalmente aqueles em que as razões para a não resposta não se ajustam a nenhuma das classificações anteriores.

Existem muitas outras classificações[1] que não interessa considerar. É contudo interessante fazer algumas considerações que identifiquem o comportamento diferenciado de certos grupos ou as consequências de certos tipos de questões (veja-se, por exemplo, Durão(1993)).

É mais fácil obter resposta nos meios rurais do que nos meios urbanos. Os agricultores passam mais tempo em casa. É mais fácil encontrar uma dona de casa do que uma pessoa empregada.

Está provado que as pessoas de mais idade apresentam uma elevada taxa de recusa a serem entrevistados, devido a factores que decorrem do

---

[1] Por exemplo, Cochran (1977) considera quatro classes: *(1) noncoverage - failure to locate or visit some units in the sample , (2) not-at-homes - persons who reside at home but are temporarily away from the house, (3) unable to answer - respondents may not have the information wanted in certain questions or may be unwilling to give it, (4) the "hard-core" - persons who adamantly refuse to be interviewed, who are incapacitated, or are far from home.*

processo de envelhecimento humano, ou a participar em sondagens feitas por telefone, devido a deficiências auditivas.

Em muitas sondagens verifica-se uma elevada taxa de não respondentes entre os homens solteiros.

Para todo o tipo de inquérito (entrevista, telefone ou correio) as pessoas com nível cultural mais elevado são, em geral, as que mais facilmente colaboram. Contrariando as previsões decorrentes deste facto verifica-se, ao tentar relacionar a taxa de recusa com a raça, que entre as pessoas brancas a não resposta é relativamente mais frequente do que entre os negros. Como diz Durão(1993), este resultado não era previsível visto que as grandes taxas de recusa se encontram entre os respondentes de nível cultural mais baixo, categoria a que em geral percentencem as pessoas de raça negra.

Como é de esperar, famílias pequenas, constituídas por uma só pessoa ou por um casal, respondem menos do que famílias numerosas. A razão encontrada baseia-se no facto de as famílias pequenas serem em grande parte constituídas por pessoas idosas mais relutantes em responder.

A falta de resposta é mais acentuada em sondagens que envolvem questões sensíveis - consumo de drogas, número de abortos efectuados,

religião - ou de carácter monetário - rendimentos pessoais e sua origem, etc.

3.3 - Quantificação da não resposta

Há muitos critérios para quantificar a não resposta. O mais simples parte da dicotomia responde - não responde.

Exemplificando: nos Quadros 3.1 e 3.2, reproduzidos graficamente nas Figuras 3.1 e 3.2, encontram-se alguns resultados do Inquérito Anual aos Museus feito pelo I.N.E. em 1993.

Quadro 3.1 - Museus, por distribuição geográfica, segundo a resposta (Continente)

Distribuição Geográfica	*Total de Museus*	*Museus que responderam*	*Taxa de resposta*
Norte	70	49	0,70
Centro	53	37	0,70
Lisboa e V.T.	129	100	0,78
Alentejo	34	22	0,65
Algarve	15	7	0,47
TOTAL	301	215	0,71

Quadro 3.2 - Museus, por natureza, segundo a resposta

(Continente, Açores e Madeira)

Natureza	*Total de Museus*	*Museus que responderam*	*Taxa de resposta*
Arte	80	63	0,79
Arqueologia	83	59	0,71
Ciências e História Natural	13	10	0,77
Ciência e Técnica	10	7	0,70
Etnografia	36	24	0,67
Museus Espec.	23	14	0,61
Outros	83	64	0,77
TOTAL	328	241	0,74

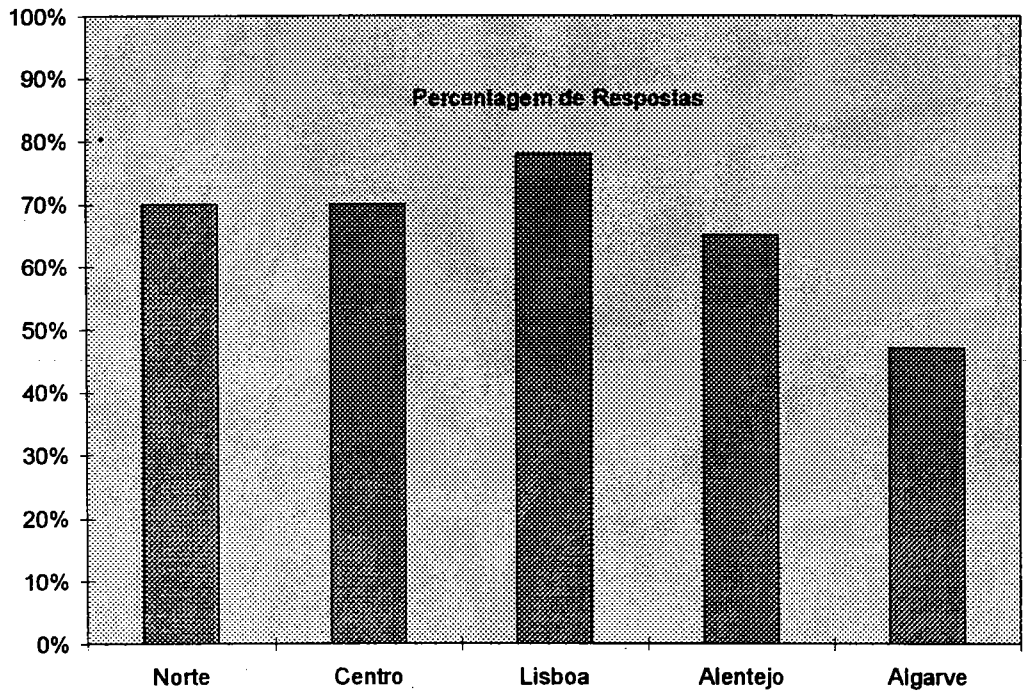


Fig.3.1

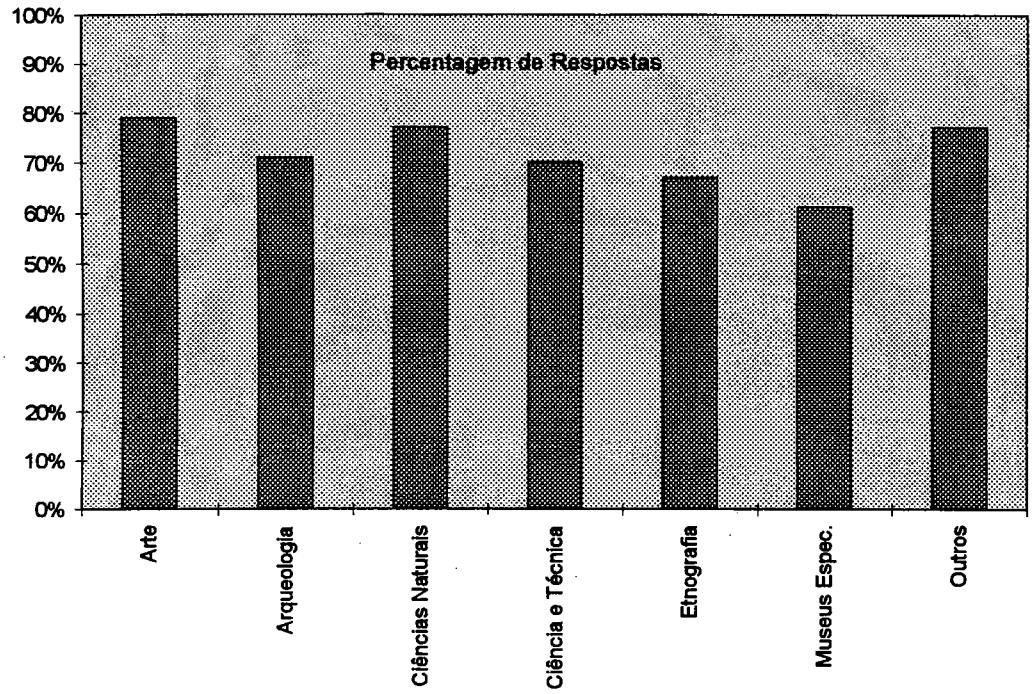


Fig.3.2

Como se verifica a taxa de resposta anda, para o conjunto, ligeiramente acima dos 70%. Pareceu interessante investigar para os dois quadros acima a homogeneidade da taxa de resposta em função da distribuição geográfica e em função da natureza dos objectos e colecções expostas. Para o efeito construiram-se duas tabelas de contingência cuja

análise forneceu os seguintes resultados:

- Distribuição geográfica dos museus:

  Qui-quadrado com 4 graus de liberdade - $\chi^2 = 8{,}028$;

  Nível de significância da hipótese da homogeneidade - $\alpha = 0{,}095$.

- Natureza dos museus:

  Qui-quadrado com 6 graus de liberdade - $\chi^2 = 4{,}821$;

  Nível de significância da hipótese da homogeneidade - $\alpha = 0{,}567$.

Se nenhum dos resultados leva a rejeição da hipótese da homogeneidade das taxas de resposta ao nível de 5%, restam poucas dúvidas de que a distribuição regional apresenta tendencialmente uma muito maior heterogeneidade.

Lessler e Kalsbeek(1992), usando a terminologia já referida, apresentam um vasto leque de medidas alternativas para a taxa de resposta. A proposta de Särndal, Swensson e Wretman(1992) parece mais interessante e mais simples. Se o inquérito envolve questões sobre p variáveis, sendo, em termos gerais, a resposta da i-ésima unidade,

$$\left(Y_{i1}, Y_{i2}, \ldots, Y_{ip}\right),$$

represente-se por,

$$r_j = \left\{i: i = 1,2, \ldots, n; Y_{ij} \Rightarrow registado\right\},$$

o subconjunto dos elementos da amostra que responderam à j-ésima

questão (para os quais Y_{ij} é conhecido). Os conjuntos,

$$r_u = \bigcup_{j=1}^{p} r_j \quad \text{e} \quad r_c = \bigcap_{j=1}^{p} r_j \ ,$$

representam, respectivamente, o subconjunto das unidades que responderam pelo menos a uma questão e o subconjunto das unidades que responderam a todas as questões. Assim, se for n a dimensão da amostra e n_u e n_c os cardinais dos subconjuntos acima, definindo $p_u = n_u / n$ e $p_c = n_c / n_u$, têm-se medidas simples de unidade de não resposta, $1 - p_u$, e de item de não resposta, $1 - p_c$.

3.4 - Efeitos de não resposta

Em qualquer sondagem a não resposta é um problema grave uma vez que raramente se consegue obter uma taxa de 100% de respostas. Consequentemente, se não é feito um esforço para compensar a falta de resposta ocorre muito provavelmente uma distorsão ou enviesamento nas estimativas que usualmente constam dos objectivos da sondagem. De certo modo esse enviesamento mede o impacto provocado pela não resposta.

Na presente secção os conceitos de unidade e item de não resposta, apesar de distintos, como foi visto anteriormente, tornam-se equivalentes

nas situações em que a sondagem envolve apenas uma questão visando determinar em cada unidade o valor de uma variável Y.

O efeito ou enviesamento causado pela não resposta vai estudar-se admitindo que se está em presença do único erro não provocado por amostragem. Suponha-se que o universo tem N unidades, $U_1, U_2, \ldots, U_N$, nas quais a variável Y assume os valores $Y_1, Y_2, \ldots, Y_N$, com média $\overline{Y}$. Suponha-se ainda que o universo se encontra dividido em dois "estratos": um, constituído por N_0 unidades, com média $\overline{Y_0}$, que não respondem se eventualmente forem inquiridas; outro, constituído por N_1 unidades, com média $\overline{Y_1}$, que respondem se eventualmente forem inquiridas. Quer dizer, na população de N unidades há N_0 potencialmente não respondentes e N_1 potencialmente respondentes. Tem-se,

$$N = N_0 + N_1 \quad \text{e} \quad N\overline{Y} = N_0\overline{Y_0} + N_1\overline{Y_1} .$$

A proporção de não respondentes é dada por,

$$W_0 = N_0 / N,$$

e a proporção de respondentes por,

$$W_1 = N_1 / N,$$

com $W_0 + W_1 = 1$. A média do universo pode então escrever-se,

(3.1) $$\overline{Y} = W_0\overline{Y_0} + W_1\overline{Y_1} .$$

A não resposta origina desde logo um decréscimo na dimensão da amostra - n - pois obtém-se n_0 não respostas e n_1 respostas,

$$n = n_0 + n_1, \qquad n_0 > 0 \Rightarrow n_1 < n .$$

No entanto, o problema não reside no decréscimo já que tomando n suficientemente grande também n_1 aumentava podendo assim trabalhar-se com uma dimensão suficiente grande para os fins em vista. O problema real reside no facto de não se dispor de informação acerca de n_0 unidades amostrais correspondentes a N_0 unidades populacionais com consequências que passam a estudar-se em diversos aspectos.

Admita-se que o objectivo é estimar a média da população,

$$\overline{Y} = \sum_{i=1}^{N} Y_i / N,$$

e que, ignorando a não resposta, se emprega como estimador a estatística que se reduz à média dos respondentes, seja,

$$\hat{\overline{Y}}_1 = \overline{y}_1 = \sum_{i=1}^{n_1} y_i / n_1 .$$

Recordando (3.1) e notando que $E_s(\hat{\overline{Y}}_1) = \overline{Y}_1$, onde $E_s(.)$ designa o valor esperado sobre o conjunto de todas as amostras possíveis, o enviesamento[1] de tal estimador vem, como facilmente se verifica,

$$(3.2) \qquad B(\hat{\overline{Y}}_1) = E_s(\hat{\overline{Y}}_1) - \overline{Y} = \overline{Y}_1 - \overline{Y} = W_0(\overline{Y}_1 - \overline{Y}_0) .$$

---

[1] Emprega-se a primeira letra da palavra anglo-saxónica *Bias* para designar o enviesamento.

Como pode ver-se, o enviesamento é o produto da proporção de não respondentes pela diferença entre as médias populacionais dos dois estratos.

Se as médias forem iguais, o que em regra não se verifica, o enviesamento anula-se. Naturalmente, também é de esperar que $W_0 \neq 0$. O enviesamento é, portanto, um facto.

Analogamente, desejando estimar-se o total da característica Y na população, seja,

$$Y_T = \sum_{i=1}^{N} Y_i ,$$

através da expressão,

$$\hat{Y}_{T1} = N\hat{\bar{Y}}_1 = N\bar{y}_1 ,$$

baseada no anterior estimador, o enviesamento devido à não resposta vem dado, depois de simples cálculos, por,

(3.3) $$B(\hat{Y}_{T1}) = E_s(\hat{Y}_{T1}) - Y_T = N_0(\overline{Y}_1 - \overline{Y}_0) ,$$

expressão que podia obter-se directamente de (3.2) e que mostra mais uma vez a importância da diferença entre as médias de respondentes e não respondentes.

Retome-se o exemplo dos Museus do Inquérito do I.N.E. que originou os Quadros 3.1 e 3.2 e considerem-se os seguintes dados acerca do pessoal ao serviço:

Quadro 3.3 - Museus: pessoal ao serviço

Total de Museus	328
Museus que responderam	241
Pessoal ao serviço dos respondentes	2683

Desejando estimar-se a média do pessoal ao serviço no conjunto total de 328 museus através da estimativa $2683/241 \approx 11,1$ baseada nos respondentes, considerando (3.2) pode conceber-se a presença de um enviesamento de ordem de:

$$(87/328)\{(2683/241) - \overline{Y_0}\} = 0,27(11,1 - \overline{Y_0});$$

desejando estimar-se o total do pessoal ao serviço, ainda no conjunto total de 328 museus, através da estimativa $328(2683/241) \approx 3652$ também baseada nos respondentes, obtinha-se por (3.3) um enviesamento da ordem de:

$$(328-241)\{(2683/241) - \overline{Y_0}\} = 87(11,1-\overline{Y_0}).$$

Evidentemente não pode calcular-se qualquer destes envisesamentos por se desconhecer $\overline{Y}_0$. Repare-se ainda que não se trata propriamente de um problema de amostragem havendo, possivemente, algo de abusivo na aplicação de (3.2) e (3.3).

Desejando estudar-se o efeito da não resposta quando se usa a variância da amostra de respondentes,

$$s_1^2 = \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2 / (n_1 - 1) \ ,$$

para estimar a variância da população,

$$S^2 = \sum_{i=1}^{N} (Y_i - \overline{Y})^2 / (N - 1) \ ,$$

repare-se, em primeiro lugar, na conhecida expressão (que atende aos dois estratos - respondentes e não respondentes - e ás respectivas dimensões, médias e variâncias),

$$(N-1)S^2 = (N_0 - 1)S_0^2 + (N_1 - 1)S_1^2 + \\ + N_0(\overline{Y}_0 - \overline{Y})^2 + N_1(\overline{Y}_1 - \overline{Y})^2. \tag{3.4}$$

O enviesamento de s_1^2 vem dado por,

$$B(s_1^2) = E_s(s_1^2) - S^2 = S_1^2 - S^2,$$

ou seja, considerando (3.1) e (3.4),

$$B(s_1^2)=\frac{N-N_1}{N-1}S_1^2-\frac{N_0-1}{N-1}S_0^2-\frac{N_0}{N-1}\left(\frac{N_1}{N}\right)^2(\overline{Y_0}-\overline{Y_1})^2+\frac{N_1}{N-1}\left(\frac{N_0}{N}\right)^2(\overline{Y_0}-\overline{Y_1})^2,$$

e, finalmente,

$$B(s_1^2)\approx W_0\left[\left(S_1^2-S_0^2\right)-W_1\left(\overline{Y}_1-\overline{Y}_0\right)^2\right]. \tag{3.5}$$

Quer dizer, o enviesamento é o produto da proporção de não respondentes e da expressão incluída entre parêntesis rectos que em geral se não anula.

3.5 - Efeitos de não resposta em casos mais complexos

Para aprofundar o estudo da não resposta e, por agora, dos seus efeitos, importa introduzir dois tipos de probabilidades.

Se, como é habitual, se admite que a escolha dos elementos que compõem a amostra é feita com uma certa aleatoriedade, toma-se π_i para designar a probabilidade de selecção do elemento U_i da população $\{U_1,U_2,...,U_N\}$.

Se o inquérito envolve apenas uma variável Y que assume na população o conjunto de valores $\{Y_1,Y_2,...,Y_N\}$ considera-se a variável

binária R_i tal que,

$$\begin{cases} R_i = 1 & se\ U_i\ responde, \\ R_i = 0 & se\ U_i\ não\ responde. \end{cases}$$

Seja ainda $p_i = \Pr ob\{R_i = 1\}$ a probabilidade de resposta da unidade U_i que se traduz pela disponibilidade do valor Y_i.

Existem duas abordagens genéricas relativas ao modo de encarar a resposta as quais se distinguem pelo modo de perspectivar a correspondente probabilidade:

(A) Abordagem **determinística** que assume que os elementos da população ou respondem $(p_i = 1)$ ou não respondem $(p_i = 0)$.

(B) Abordagem **estocástica** que assume para todos os elementos da população, $0 \leq p_i \leq 1$.

Estas duas abordagens estão relacionadas, sendo a primeira evidentemente um caso particular da segunda.

Para analisar o efeito da não resposta considere-se o problema estudado por Platek, Singh e Tremblay(1977) e que consiste em investigar o enviesamento de um estimador do total da população,

$$Y_T = \sum_{i=1}^{N} Y_i \ ,$$

considerando, em primeiro lugar, a localização e a solicitação das unidades para obtenção de resposta na sondagem e, em segundo lugar, a recolha dos dados. Caso não se obtenha resposta da i-ésima unidade, o valor desconhecido Y_i é substituído por um valor artificial Z_i naturalmente sujeito a erro. Nestas circunstâncias o valor associado com a i-ésima unidade da população é dado por,

$$\hat{Y}_i = R_i Y_i + (1 - R_i) Z_i \ ,$$

e o estimador do total de Platek, Singh e Tremblay(1977) assume a forma,

(3.6) $$\hat{Y}_{PST} = \sum_{i=1}^{N} \hat{Y}_i \ .$$

Para estabelecer as propriedades do estimador de Platek, Singh e Tremblay convem introduzir algumas premissas e considerandos.

Primeiro, supõe-se que,

$$Z_i = Y_i - \varepsilon_i ,$$

onde ε_i representa o erro cometido na substituição operada no respectivo não respondente, que os $\varepsilon_i , i = 1,2,...,N$ são mutuamente independentes entre os membros da população e que,

$$E_Z(\varepsilon_i) = B_i \quad , \quad Var_Z(\varepsilon_i) = \sigma_i^2 \ ,$$

onde $E_Z(.)$ e $Var_Z(.)$ designam o valor esperado e a variância considerando todas as possíveis repetições do processo de substituição.

Segundo, considera-se que,

$$E_r(R_i) = p_i\,, \quad Var_r(R_i) = p_i(1-p_i)\,,$$

onde $E_r(.)$ e $Var_r(.)$ representam o valor esperado e a variância segundo as repetidas aplicações do processo de localização e solicitação e assume-se igualmente a independência mutua de R_i e R_j $(i \neq j)$.

Nestes termos não é difícil mostrar, depois de alguns cálculos, que o estimador de Platek, Singh e Tremblay tem erro quadrático médio dado pela expressão,

$$\begin{aligned} EQM(\hat{Y}_{PST}) &= Var(\hat{Y}_{PST}) + B^2(\hat{Y}_{PST}) \\ &= \left[\sum_{i=1}^{N} p_i(1-p_i)B_i^2 + \sum_{i=1}^{N}(1-p_i)\sigma_i^2\right] + \left[\sum_{i=1}^{N}(1-p_i)B_i\right]^2 \end{aligned} \tag{3.7}$$

em que,

$$\sum_{i=1}^{N} p_i(1-p_i)B_i^2\,, \tag{3.8}$$

representa a componente da variância atribuída à natureza estocástica da não resposta e,

$$\sum_{i=1}^{N}(1-p_i)\sigma_i^2\,, \tag{3.9}$$

a componente da variância atribuída à variação estocástica dos Z_i. O enviesamento,

$$\sum_{i=1}^{N}(1-p_i)B_i\,, \tag{3.10}$$

depende ao mesmo tempo da probabilidade de resposta e da centragem da imputação.

A análise da expressão (3.7) para certas situações é particularmente elucidativa.

Mantendo a abordagem estocástica pode notar-se o seguinte aspecto: quando não se procede a qualquer ajustamento, isto é, quando se toma por exemplo $Z_i = 0, i = 1,2,...,N$, sai,

$$Var(\hat{Y}_{PST}) = \sum_{i=1}^{N} p_i(1-p_i)Y_i^2 , \quad B(\hat{Y}_{PST}) = -\sum_{i=1}^{N}(1-p_i)Y_i .$$

Passando a perfilhar a abordagem determinística em que $p_i = 0$ para os N_0 não respondentes e $p_i = 1$ para os N_1 respondentes e supondo mais uma vez que não se procede a qualquer ajustamento, sai,

$$Var(\hat{Y}_{PST}) = 0, \quad B(\hat{Y}_{PST}) = -\sum_{i=1}^{N_0} Y_i .$$

Mais interessante é destacar, no quadro da mesma abordagem, as consequências, quanto ao enviesamento, do ajustamento ou da substituição

operada sobre os não respondentes, consequências que se traduzem por,

$$Var(\hat{\bar{Y}}_{PST}) = \sum_{i=1}^{N_0} \sigma_i^2 , \quad B(\hat{\bar{Y}}_{PST}) = -\sum_{i=1}^{N_0} B_i = N_0(\bar{Z}_0 - \bar{Y}_0) ,$$

onde,

$$\bar{Z}_0 = \sum_{i=1}^{N_0} Z_i / N_0 ,$$

resultado que mostra: (1) o desaparecimento dos factores devidos à natureza estocástica da não resposta; (2) a boa estratégia de substituição que consiste, pensando ainda no enviesamento, em proceder de modo a obter, quando possível, a igualdade entre a média dos substitutos e dos substituídos. Em particular, se o valor dos Z_i for igual à média dos respondentes - $Z_i = \bar{Y}_1 , i = 1,2,\ldots,N$ - tem-se, como era de esperar, a expressão (3.3),

$$B(\hat{\bar{Y}}_{PST}) = N_0(\bar{Y}_1 - \bar{Y}_0) .$$

Os efeitos da falta de resposta no enviesamento de estimadores da média populacional são de certo modo paralelos aos efeitos verificado em estimadores do total e não interessa que sejam aqui retomados.

Até agora todas as considerações são baseadas na aplicação da chamada inferência clássica[1] nos problemas de amostragem de populações finitas. Em que consiste a inferência clássica no contexto da amostragem de populações finitas e quais são as opções não clássicas?

Para começar, considerando apenas uma característica relevante Y, a inferência clássica considera como base de raciocínio que a população de N valores, seja $(Y_1, Y_2, ..., Y_N)$, constitui um **vector constante.** A inferência baseada em modelos[2] é mais recente e possivelmente aquela que conta mais intensa investigação. Considera que a população de valores, que de acordo com a prática habitual convem representar agora por variáveis minúsculas $(y_1, y_2, ..., y_N)$, é uma realização do **vector aleatório** $(Y_1, Y_2, ..., Y_N)$ gerada por um processo estocástico, isto é, pode considerar-se uma amostra de uma **super-população**.

A inferência clássica ou "*design based*" e a "*model-based*" distinguem-se pelo tipo de suposições que avançam ao lidar com o problema da falta de resposta, estando longe de arrumada a controvérsia sobre qual a melhor "filosofia" para tratar tal problema.

---

[1] Em nosso entender a designação de inferência clássica no presente contexto não é muito própria uma vez que a mesma se aplica usualmente a uma corrente, no quadro mais vasto da inferência estatística, que se opõe, por exemplo, à inferência Bayesiana.

[2] Veja-se o excelente artigo de Royall(1992).

A primeira conta com o mecanismo aleatório ("*randomization*") utilizado para seleccionar a amostra do conjunto fixo $(Y_1, Y_2, ..., Y_N)$, processo que faz acompanhar de hipóteses sobre a relação entre respondentes e não respondentes[1].

A segunda, em contraste, trata a não resposta considerando $(Y_1, Y_2, ..., Y_N)$ como um vector aleatório e insistindo em que as inferências relevantes são mais do tipo preditivo.

Quer dizer, dando pouca ou nenhuma atenção ao esquema de recolha da amostra[2] ou à forma como a não resposta reduz de n para n_1 a dimensão da amostra disponível, a inferência baseada em modelos sustenta, em última análise, que o que interessa é **predizer**[3] os valores associados com os não respondentes e com os elementos da população que não foram seleccionados para fazer parte da amostra.

Não é propósito do presente trabalho aprofundar a inferência baseada em modelos. No entanto vai fazer-se-lhe uma breve referência tendo

---

[1] Por exemplo, como vai ver-se adiante, admitindo que em certas subclasses as médias dos respondentes e dos não respondentes são iguais.

[2] Que interessa considerar **apenas** como meio de evitar recolhas enviesadas, mas em que as probabilidades de selecção pouco interessam.

[3] Sobre diferença essencial entre **estimar** parâmetros não observáveis e **predizer** valores de variáveis obseváveis pode ler-se Geisser(1993).

presente que a ideia base continua a ser o estudo dos efeitos da não resposta.

Considerando o vector aleatório $(Y_1,Y_2,...,Y_N)$ suponha-se que Y_i é independente de Y_j para todo o $i \neq j$ e tome-se o modelo linear,

$$\begin{aligned} & Y_i = \beta X_i + \xi_i, \\ & E_m(\xi_i) = 0, \quad Var_m(\xi_i) = \sigma^2 X_i, \\ & i = 1,2,...,N, \end{aligned} \tag{3.11}$$

onde β e σ^2 são parâmetros, $E_m(.), Var_m(.)$ representam valor esperado e variância em relação ao modelo e $(X_1,X_2,...,X_N)$ são valores da variável auxiliar X que se supõe conhecida para todos os elementos da população (por exemplo, sexo, região, cor da pele, etc.). A validade do modelo (3.11), ou de qualquer outro modelo empregue, tem sempre de ser questionada porquanto condiciona a qualidade dos resultados obtidos.

O modelo (3.11) pressupõe que existe tendência para uma certa proporcionalidade entre os Y_i e os X_i e estabelece um vínculo entre o conjunto dos Y_i observados (respondentes na amostra) e os Y_i restantes ou não observados (não respondentes ou não incluídos na amostra), conjuntos que passam a designar-se respectivamente por {o} e {r}.

O objectivo é estimar o total da população,

(3.12) $$Y_T = \sum_{i\in\{o\}} y_i + \sum_{i\in\{r\}} Y_i \ ,$$

problema que é equivalente a predizer o valor da variável aleatória,

$$\sum_{i\in\{r\}} Y_i \ .$$

O estimador proposto por Royall(1992),

(3.13) $$\hat{Y}_R = \sum_{i\in\{o\}} y_i + \hat{\beta} \sum_{i\in\{r\}} X_i \quad \text{com} \quad \hat{\beta} = \sum_{i\in\{o\}} y_i \,/\, \sum_{i\in\{o\}} X_i \ ,$$

é, suposta a validade de (3.11), o estimador BLU de (3.12) porque,

$$\hat{\beta} \sum_{i\in\{r\}} X_i$$

é o estimador BLU de,

$$\sum_{i\in\{r\}} Y_i \ .$$

Note-se que o estimador (3.13) é afinal o bem conhecido estimador rácio,

$$\hat{Y}_R = \sum_{i=1}^{N} X_i \left(\sum_{i\in\{o\}} y_i \,/\, \sum_{i\in\{o\}} X_i \right),$$

que como acaba de ver-se é não enviesado (por ser BLU) e insensível[1] ao esquema de amostragem e à presença de não respondentes.

Recorde-se que a propriedade do não enviesamento no quadro de um modelo ("*model unbiased*") é por vezes criticada pelo facto de estimadores não enviesados para certo modelo deixarem de o ser para outros modelos. Semelhante facto levanta a questão da robustez, isto é, a manutenção de certas propriedades perante modificações, não muito drásticas, na estruturação do modelo. No entanto, não põe em causa a importância que tem ter-se encontrado, quando o modelo é válido, insista-se, uma solução para o problema da não resposta.

---

[1] Lessler e Kalsbeek(1992), ao referirem-se às inferências baseadas em modelos, do tipo aqui considerado, expressam-se de forma que merece ser reproduzida: *"Inference... might be based exclusively on this model and data from n_1 respondents, with limited concern as to how the original sample of size n was selected or how nonresponse reduced to n_1 the number of sample observations."*.

CAPÍTULO 4

MÉTODOS PARA REDUÇÃO DO EFEITO DA NÃO RESPOSTA

Analisando a expressão (3.2) pode ver-se, por exemplo, na estimação da média da população, que o enviesamento devido à não resposta é independente do número de unidades recolhidas com sucesso na amostra. Deste modo o enviesamento não pode ser reduzido por um aumento da dimensão da amostra, devendo recorrer-se, para o efeito, a outros métodos, um dos quais consiste em reduzir a proporção de potenciais não respondentes ($W_0 = N_0 / N$) - os N_0 potenciais não respondentes constituem um grupo bastante heterógeneo - uma vez que pouco se pode fazer acerca da diferença: $(\overline{Y}_1 - \overline{Y}_0)$. Que processos podem então contribuir para essa redução?

Existem dois processos fundamentais:

(1) Preventivo;

(2) Compensação da falta de resposta.

Antes e durante a recolha de dados são tomadas medidas efectivas para reduzir a falta de resposta a níveis insignificantes de modo que alguma

falta de resposta que persista cause pouca ou quase nenhuma perturbação à validade das inferências.

Na presente fase do trabalho vão apresentar-se técnicas que conservam a taxa de resposta alta, tais como insistências ou sucessivas abordagens aos elementos seleccionados para a amostra e que não respondem à primeira. Este método pode utilizar-se em inquéritos pessoais porta a porta, entrevistas pelo telefone ou em questionários enviados pelo correio.

4.1 - Métodos preventivos da falta de resposta

Os métodos preventivos são vários não sendo possível fazer uma lista exaustiva. A prevenção começa logo à partida quando se faz o planeamento da forma como deve decorrer a sondagem. Estes métodos baseiam-se, por um lado, no encorajamento à obtenção de resposta e, por outro lado, na "insistência" até à obtenção da mesma, mas tal não significa que em qualquer das situações se obtenha um aumento significativo de respondentes.

Como já foi dito, os não respondentes constituem um grupo bastante heterógeneo, mas principalmente o tipo de não respondentes varia de acordo com o modo como é feita a sondagem, isto é, consoante a sondagem é feita por entrevista directa, por telefone ou por carta e também

consoante o tema da sondagem. Como diz Lemeshow(1985): "*The methods used to encourage response should be carefully and specifically tailored to particular nonrespondents, and the type of survey being used*".

Nas sondagens ou inquéritos porta a porta ou pelo telefone, o habitante, ou a pessoa em particular com quem pretende falar-se, pode não se encontrar em casa, estar doente ou, por qualquer outro motivo, estar indisponível ou incapacitado para falar. Pode ainda ter de interromper a sua colaboração, como por exemplo se é feito um questionário demasiado longo numa hora inconveniente para o entrevistado. O procedimento recomendado é tentar entrar em contacto em outras ocasiões mais convenientes, utilizando, talvez, entrevistadores com mais habilidade para fazer falar os entrevistados. Deve proceder-se de modo semelhante em sondagens por telefone. Consegue-se assim, em geral, diminuir substancialmente a não resposta.

Por vezes, se é utilizado um questionário por correspondência, há necessidade em proceder de seguida a implacáveis insistências, quer por correspondência, quer por telefone, quer ainda, por entrevista pessoal.

Para exemplificar uma situação de sondagem por correspondência vai utilizar-se o Quadro 4.1:

Quadro 4.1 - Sondagem por correspondência no Norte da Carolina a Cultivadores de árvores de fruto[1]

(número total: 3116)

	Número de envios			Total
	1	2	3	Não resposta	Respostas e não respostas
% de Respostas	10	17	14	59	100
Média de árvores de fruto por quinta	456	386	340	290	329

Foram feitos três envios sucessivos aos agricultores - em que os dados estavam disponíveis para a população em relação à questão "número de árvores de fruto".

Observando o Quadro 4.1 verifica-se que os agricultores que possuem menor número de árvores de fruto têm mais relutância em nos responder.

[1] Fonte: Finkner(1950) referido por Cochran (1977)

Para exemplificar o que se passa com uma sondagem telefónica vai utilizar-se uma situação concreta referida por Selbold(1988), segundo a qual, as sondagens telefónicas deparam com dois problemas administrativos distintos - que particularmente as diferenciam de outros tipos de sondagens que não sejam por correio - pois têm "falta" de contacto visual e físico com as unidades amostrais.

Este problema não se coloca nas sondagens "*face to face*" as quais não exigem alterações ao modo de prosseguir a sondagem, até porque os técnicos quando contactam pela primeira vez uma morada seleccionada obtêm frequentemente informação acerca de quando poderão voltar a contactar essa mesma morada.

Com as sondagens telefónicas a informação que se obtém dos números que "não atendem" é praticamente nula embora tenha a vantagem de não ser muito dispendioso.

A ilustração que se encontra representada no Quadro 4.2 mostra a evolução da percentagem de resposta, ao longo do tempo, de sucessivas chamadas telefónicas, e faz parte de uma pesquisa metodológica em entrevistas por telefone conduzidas pelo *U.S. Census Bureau*, inserida no "*U.S. National Crime Survey*".

Quadro 4.2 - Número de chamadas telefónicas para completar uma entrevista

Número de tentativas	Entrevistas conseguidas	Percentagem	Percentagem acumulada
1	369	9,2	9,2
2	494	12,3	21,4
3	406	10,1	31,5
4	393	9,8	41,3
5	261	6,5	47,8
6	228	5,7	53,4
7	216	5,4	58,8
8	201	5,0	63,8
9	161	4,0	67,8
10	134	3,3	71,1
11	119	3,0	74,1
12	114	2,8	76,9
13	80	2,0	78,9
14	66	1,6	80,5
15	53	1,3	81,8
16	65	1,6	83,4
17	54	1,3	84,8
18	40	1,0	85,8
19	48	1,2	87,0
20	25	0,6	87,6
21	31	0,8	88,4
22	29	0,7	89,1
23	33	0,8	89,9
24	18	0,4	90,3
25	27	0,7	91,0
25+	362	9,0	100,0

Nesta investigação foi utilizada uma metodologia diferente do comum para obter parte dos dados pretendidos pelo que parece de interesse fazer referência a este trabalho. Como já foi dito é relativamente escassa a informação que se obtem dos "números de telefone que não atendem", sendo necessário por vezes efectuar repetidas chamadas telefónicas; este problema é exacerbado no esquema de amostragem "*two stage ramdom digit dialing* (RDD) ". Este esquema, quando um número é seleccionado para a amostra, depois de alguns dias de chamadas sem atender é designado por não residêncial, sendo então outro número seleccionado para substituir aquele na amostra, recebendo assim o número de substituição a primeira chamada telefónica depois de ter decorrido algum tempo sobre o início da sondagem.

Note-se no Quadro 4.2 que 21,4% das entrevistas foram obtidas em duas chamadas e 53,4% ao fim de seis chamadas.

Os métodos de compensação reduzem substancialmente a falta de resposta (quer as unidades, quer os itens de não resposta); no entanto, não a eliminam por completo. Assim é necessário recorrer a outros métodos, nomeadamente os já referidos "**métodos de compensação da falta de resposta**"

A grande maioria destes métodos pode ser aplicada, quer às unidades não respondentes, quer aos itens de não resposta, embora alguns sejam "demasiado pesados" para serem aplicados aos itens.

4.2 - Compensação da falta de resposta

Passam a descrever-se algumas técnicas para permitir a estimação não enviesada ou menos enviesada.

Substituição de unidades não respondentes

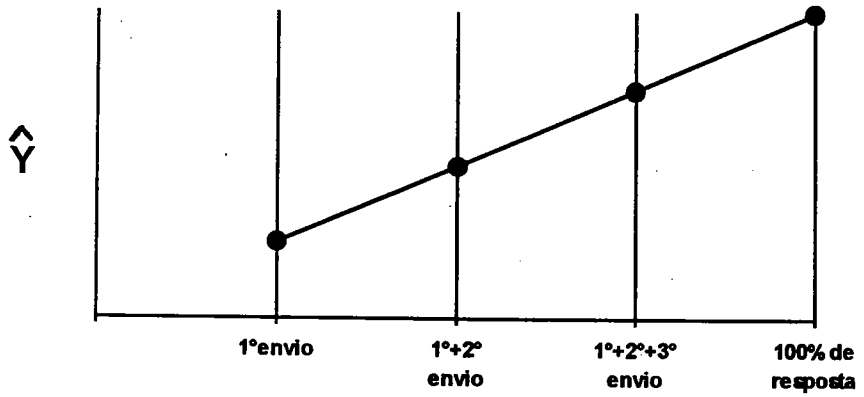
Este método consiste em substituir, na amostra inicial, unidades não respondentes por outras unidades pertencentes à população - frequentemente uma unidade não incluída na amostra inicial.

A escolha do substituto (respondente) pode ser ou não realizada de modo aleatório e o objectivo é substituir cada não respondente por outro que seja semelhante a esse em relação à maioria das variáveis em estudo. De facto, o método de substituição reduz o enviesamento devido à não resposta desde que os substitutos sejam colectivamente idênticos àqueles que substituem. Contudo, não elimina por completo o enviesamento (devido à falta de resposta), e justifica, portanto, o esforço que em regra se faz, antes de utilizar este método, no sentido de obter resposta dos não respondentes da amostra inicial.

Método de extrapolação

Este método, apesar de incluído nos métodos de compensação, é de certo modo uma composição de um método preventivo com um método de compensação propriamente dito. É utilizado principalmente nas sondagens por correspondência; após o envio do primeiro questionário é enviado um segundo aos não respondentes, repetindo-se sucessivamente a operação até se atingir o número de envios máximo previamente fixado (no plano da sondagem).

Fig. 4.1



Utilizando os resultados dos envios sucessivos, produzem-se os resultados pretendidos por extrapolação como no exemplo da Fig. 4.1, em que se consideram três envios onde a "100% de resposta" correspondem os valores obtidos por extrapolação para o total da população representado no eixo das ordenadas.

Como refere Grosbras(1987), deve assegurar-se a coerência das extrapolações.

Subamostragem dos não respondentes

Outro método de compensação parte, como muitos outros, da convicção de que o problema do enviesamento por falta de resposta pode ser resolvido ou, mais realisticamente, ser bastante reduzido em muitas situações. A ideia mestra é utilizar na estimação os respondentes da amostra e os (muito desejavelmente) respondentes de uma subamostra aleatória retirada dos não respondentes. Os resultados do método dependem, dramaticamente, do sucesso no esforço novamente feito para fazer falar os inicialmente não respondentes.

Existem vários esquemas de subamostragem dos não respondentes, ideia desenvolvida inicialmente por Hansen e Hurwitz(1946), que pode descrever-se em termos gerais utilizando apenas uma variável - Y - e um esquema de amostragem arbitrário para obter a amostra inicial e posteriormente a subamostra. Um excelente tratamento encontra-se em Särndal, Swesson e Wretman(1992); no entanto vai aqui seguir-se o tratamento mais elementar de Cochran(1977) e Grosbras(1987)[2]. O objectivo é estimar a média da população.

[2] Veja-se também Durão(1993).

Supondo a população dividida em dois estratos na proporção de $W_0 = N_0 / N$ não respondentes e $W_1 = N_1 / N$ respondentes, considere-se uma primeira amostra de dimensão n na qual há n_0 não respondentes e n_1 respondentes. Seguidamente dos n_0 não respondentes tira-se uma subamostra de dimensão n_0^*, adoptando-se em ambos os casos amostragem casual simples. O estimador de Hansen-Hurwitz pode então escrever-se,

$$\hat{\bar{Y}}_{HH} = \frac{n_0}{n}\bar{y}_0^* + \frac{n_1}{n}\bar{y}_1, \tag{4.1}$$

em notação facilmente compreensível: $\bar{y}_0^*$ é a média da subamostra dos não respondentes , $\bar{y}_1$ é a média dos respondentes.

Para mostrar que o estimador $\hat{\bar{Y}}_{HH}$ é não enviesado calcula-se o valor esperado em duas fases considerando primeiro n_0 / n fixo (que implica n_1 / n evidentemente fixo) e considerando depois a variabilidade em n_0 / n. Na primeira fase, tem-se, evidentemente,

$$E(\bar{y}_0^* | n_0 / n) = \bar{Y}_0 \quad , \quad E(\bar{y}_1 | n_0 / n) = \bar{Y}_1 \ ,$$

donde,

$$E(\hat{\bar{Y}}_{HH} | n_0 / n) = \frac{n_0}{n}\bar{Y}_0 + \frac{n_1}{n}\bar{Y}_1 \ ,$$

e finalmente,

$$E(n_0 / n) = N_0 / N \, , E(n_1 / n) = N_1 / N \;\; \Rightarrow \;\; E(\hat{\bar{Y}}_{HH}) = \bar{Y} \; .$$

A variância obtém-se através de um Teorema de Rao (vejam-se Cochran(1977) ou Grobras(1987); o segundo apresenta expressões que a permitem estimar),

$$Var(\hat{\bar{Y}}_{HH}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 + \frac{1}{n} \frac{N_0}{N} \left(\frac{1-f}{f}\right) S_0^2 \; ,$$

onde $f = n_0^* / n_0$, S_0^2 é a variância do estrato populacional dos não respondentes e S^2 a variância da população. É interessante notar que a primeira parcela do 2º membro é a bem conhecida expressão da variância da média quando só há respondentes e a segunda parcela resulta de haver não respondentes e é função de f e tanto maior quanto menor for f.

Método da resposta casualizada

Inquéritos que lidam com questões delicadas como consumo de droga, abortos, declaração de rendimentos e mais actualmente com o grave problema da SIDA, colocam problemas particulares de não resposta conforme, aliás, foi referido. Em tais inquéritos acontece que muitos

elementos da amostra seleccionada ou se recusam a participar ou dão respostas falsas ou evasivas.

Curiosamente existem para os casos acima referidos métodos que, sem deixar de ser informativos, protegem os respondentes permitindo-lhes manter praticamente o anonimato. São os métodos baseados na **técnica da resposta casualizada** ("*randomized response technique*") em que os dados são obtidos por entrevistas pessoais de tipo especial que visam garantir a cooperação dos indivíduos seleccionados. Estes métodos são frequentemente comparados com a recolha de dados através de questionários anónimos, já que o inquirido em vez de dar resposta à questão sensível "fornece" ao analista dados inócuos que no entanto permitem mais tarde obter estimativas válidas para a população.

Suponha-se, por hipótese, que se requer uma estimativa para a presença ou proporção de determinado atributo sensível "***K***" nos elementos da população. A proposta original de Warner(1965) assegura o anonimato do seguinte modo: a todos os elementos de uma amostra casual é pedido para fazer uma escolha através de um processo aleatório que pode ser tão simples como o lançamento de um dado ou a tiragem de uma carta de um baralho. A escolha aleatória - cujo resultado o inquirido guarda para si - é

entre duas situações:

⇒ (A) Tenho o atributo "***K***"?

(B) Não tenho o atributo "***K***"?

Admite-se que o sorteio feito pelo inquirido atribui a (A) a probabilidade *p* e a (B) a probabilidade *1-p* (sempre com $p \neq 1/2$, por exemplo, $p=1/3$ se, ao lançar um dado regular, a situação (A) é escolhida quando saem 1 ou 2 pontos e $1-p=2/3$ se a situação (B) é escolhida quando saem 3, 4, 5 ou 6 pontos). Feito o sorteio e obtido o resultado - que só o inquirido conhece - o entrevistado tem de responder, com absoluta honestidade,

Verdadeiro=1; Falso=0.

Isto é, o indivíduo fornece apenas a resposta - verdadeiro ou falso - mas mais ninguém sabe qual a questão a que respondeu. Tal como refere Crespo(1976) deveria dar-se a este método o nome de pergunta casualizada em vez de resposta casualizada já que o que de facto acaba por ser aleatório é a escolha da questão.

Os valores obtidos por este método são, portanto, binários, quer dizer, tem-se $X_i = 0$ (Falso) ou $X_i = 1$ (Verdadeiro) para $i=1,2,...,n$. Além disso,

uma simples regra do cálculo de probabilidades permite escrever,

$$\begin{cases} P(X_i = 1) = p\phi + (1-p)(1-\phi), \\ P(X_i = 0) = p(1-\phi) + (1-p)\phi, \end{cases}$$

onde ϕ é a proporção de elementos da população que possuem o atributo "***K***" e que pretende estimar-se. Fazendo,

$$\omega = P(X_i = 1) = 1 - p + \phi(2p-1) \ ,$$

tem-se que o estimador da máxima verosimilhança de ω, seja $\hat{\omega}$, é dado pela proporção de valores 1 na amostra; consequentemente, pelo princípio da invariância,

$$\hat{\phi} = \frac{\hat{\omega} + p - 1}{2p-1}, \quad p \neq \tfrac{1}{2} \ ,$$

é o estimador da máxima verosimilhança de ϕ que facilmente se verifica não ser enviesado e ter variância,

$$Var(\hat{\phi}) = \frac{p(1-p)}{(2p-1)^2}, \quad p \neq \frac{1}{2} \ .$$

Introduzindo os "pseudo-valores",

$$\hat{Y}_i = \frac{X_i + p - 1}{2p-1}, \quad p \neq \tfrac{1}{2} \ ,$$

e partindo da hipótese de que a selecção da amostra é feita utilizando um esquema aleatório em que a unidade $U_i, i = 1,2,...,N$, é escolhida com probabilidade π_i, tem-se, seguindo Särndal, Swesson e Wretman(1992), que o estimador do número total de indivíduos com o atributo "***K***" na população $\left(\sum_{i=1}^{N} Y_i\right)$, dado por,

$$\hat{Y}_{SSW} = \sum_{i=1}^{n} \frac{\hat{Y}_i}{\pi_i},$$

é não enviesado.

Deve notar-se que os procedimentos preventivos, assim como os três métodos de compensação já expostos, são aplicados durante o período de recolha dos dados; os métodos de compensação que vão ser agora desenvolvidos são utilizados após a recolha dos dados.

Ajustamentos por ponderação

Hipóteses acerca do mecanismo de resposta e acerca das relações entre variáveis são introduzidas, como vai ver-se, para construir estimadores que sofrem ajustamentos para neutralizar dentro do possível a falta de resposta.

Estes métodos (clássicos) lidam com o problema da não resposta ajustando os dados obtidos dos respondentes por forma a serem utilizados para a estimação "compensando" de certo modo a falta de resposta. De facto o ajustamento é feito através de um factor numérico ponderador instituído para o efeito.

Quando pretende estimar-se o total Y_T da população a partir de uma amostra probabilística de dimensão n, uma boa solução consiste em empregar o estimador de Horvitz-Thompson,

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \pi_i^{-1} Y_i \ . \tag{4.2}$$

Supõe-se que a unidade $U_i, i = 1,2,...,N$, é escolhida com probabilidade π_i, sendo então π_i^{-1} o factor de ponderação.

O estimador não enviesado (4.2) pode apenas ser calculado na ausência de não resposta e tem variância dada pela expressão,

$$Var(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} Y_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j \ .$$

em que π_{ij} é a probabilidade para que as unidades U_i e U_j pertençam ambas à amostra. Mas, na presença de não resposta, o estimador (4.2) é

enviesado. Assim tem de utilizar-se um "ponderador revisto",

$$(\pi_i p_i)^{-1} ,$$

que tenha em atenção, quer a probabilidade de selecção π_i, quer a probabilidade de resposta p_i (que deve ser superior a zero, para todo o $i=1,...,N$). O estimador de Horvitz-Thompson modificado passa a ser,

(4.3) $$\hat{Y}^*_{HT} = \sum_{i=1}^{n_1} (\pi_i p_i)^{-1} Y_i \ .$$

Uma vez que p_i é desconhecido deve ser estimado e é o modo como é estimado que distingue os vários métodos de ajustamento alguns dos quais vão ser expostos a seguir.

Ajustamento de Politz-Simons

Baseia-se num método proposto em 1946 por Hartley e mais tarde - 1949/50 - desenvolvido por Politz e Simons. Consiste em realizar apenas uma visita e ponderar os resultados obtidos com a probabilidade de encontrar o entrevistado nessa única visita. Uma pessoa que está em casa uma proporção de tempo p durante o período em que se realizam entrevistas, tem probabilidade p de ser encontrada em casa quando visitada. O método de Politz-Simons pressupõe que as visitas se realizam durante seis dias de cada semana, perguntando-se ao entrevistado colaborante se se encontrava em casa em cada um dos cinco dias

anteriores. Dado que há seis casos possíveis, $t = 0,1,2,3,4,5$, toma-se para estimador de p,

$$\hat{p} = (1+t)/6 \ ,$$

donde sai o estimador de Politz-Simons para a média,

$$\hat{\bar{Y}}_{PS} = \frac{\sum_{t=0}^{5} \frac{6n_t \overline{Y_t}}{1+t}}{\sum_{t=0}^{5} \frac{6n_t}{1+t}} \ , \qquad (4.4)$$

onde n_t é o número de unidades que disseram ter estado em casa em t dos dias anteriores e $\overline{Y_t}$ a respectiva média. Este estimador é menos enviesado do que a média simples dos respondentes mas a sua variância (que tem expressão algo complicada - veja-se Cochran(1977)) é maior porquanto as ponderações são estimadas.

Ajustamento por ponderação em classes

Este método estima as probabilidades de resposta em H subconjuntos disjuntos que passam a designar-se por células de ajustamento. Para uma orientação na formação das células nada parece melhor do que a recomendação de Lessler e Kalsbeek(1992): "*The ideal set of variables would be those strongly associated with the major study variables but mutualy unrelated. It also seems preferable to define cells by a coarse division on several acceptable variables than to form the same number of*

cells by a finer division on the best single variable among those considered acceptable."

Para fazer uma ligeira abordagem do método suponha-se que o propósito é estimar a média da população. Considere-se, para facilitar a exposição, que a população está dividida em H estratos[1],

$$N = N_1 + N_2 + ... + N_H \ ,$$

correspondentes às H células consideradas na amostra de dimensão,

$$n = n_1 + n_2 + ... + n_H \ .$$

Se π_{hi} designa a probabilidade de escolha da i-ésima unidade da h-ésima célula de ajustamento, π_{hi}^{-1} é o respectivo peso amostral não ajustado. Faça-se $\varpi_{hi} = 1/\pi_{hi}$.

Se for ,

$$\overline{Y}_{1h} = \frac{\sum_{i=1}^{N_{1h}} Y_{hi}}{N_{1h}} \ , \tag{4.5}$$

a média dos potenciais respondentes na h-ésima célula, o respectivo estimador é dado por,

$$\hat{\overline{Y}}_{1h} = \frac{\sum_{i=1}^{n_{1h}} \varpi_{hi} Y_{hi}}{\sum_{i=1}^{n_{1h}} \varpi_{hi}} \ , \tag{4.6}$$

[1] Para uma análise mais profunda, em que o número de células pode variar de amostra para amostra, veja-se Särndal, Swesson e Wretman(1992).

onde n_{1h} é o número de unidades amostrais respondentes na h-ésima célula ($n_{0h} = n_h - n_{1h}$ é o número de unidades amostrais não respondentes na mesma célula). Para a média da população, tem-se o estimador (com ponderação em classes),

$$\hat{\bar{Y}}_{PC} = \sum_{h=1}^{H} \hat{\Delta}_h \hat{\bar{Y}}_{1h} \quad . \tag{4.7}$$

Agora, na expressão anterior, $\hat{\Delta}_h$ estima N_h / N e tem, portanto, a forma,

$$\hat{\Delta}_h = \sum_{i=1}^{n_h} \varpi_{hi} \Big/ \sum_{h=1}^{H} \sum_{i=1}^{n_h} \varpi_{hi} \; ; \tag{4.8}$$

Fazendo a substituição de (4.6) e (4.8) em (4.7) obtem-se a forma final do estimador (4.7)

$$\hat{\bar{Y}}_{PC} = \sum_{h=1}^{H} \sum_{i=1}^{n_{1h}} \varpi^*_{hi} Y_{hi} \Big/ \sum_{h=1}^{H} \sum_{i=1}^{n_{1h}} \varpi^*_{hi} \; , \tag{4.9}$$

com,

$$\varpi^*_{hi} = \varpi_{hi} \left(\sum_{i=1}^{n_h} \varpi_{hi} \Big/ \sum_{i=1}^{n_{1h}} \varpi_{hi} \right) ;$$

Note-se que ϖ_{hi}^{*} é ajustado da não resposta porquanto $\hat{p}_{hi}$, o estimador de p_{hi}, probabilidade de resposta da unidade i-ésima da h-ésima célula (igual para todos os elementos da h-ésima célula), é dado por,

$$\hat{p}_{hi} = \frac{\sum_{i=1}^{n_{1h}} (\pi_{hi})^{-1}}{\sum_{i=1}^{n_h} (\pi_{hi})^{-1}} = \frac{\sum_{i=1}^{n_{1h}} \varpi_{hi}}{\sum_{i=1}^{n_h} \varpi_{hi}}. \tag{4.10}$$

Kalton(1983), na esteira de Thomsen(1973), deduziu para o estimador (4.7) a expressão do enviesamento,

$$B(\hat{\bar{Y}}_{PC}) = \sum_{h=1}^{H} \Delta_h \frac{N_{0h}}{N} \left(\bar{Y}_{1h} - \bar{Y}_{0h} \right), \tag{4.11}$$

e da variância que aqui não vai reproduzir-se (veja-se Lessler e Kalsbeek(1992)). Da expressão do enviesamento retira-se importante conclusão sobre as vantagens de considerar células homogéneas, nomeadamente células em que sejam iguais as médias dos respondentes ($\bar{Y}_{1h}$) e dos não respondentes ($\bar{Y}_{0h}$). A expressão da variância permite fazer a comparação do estimador (4.7) com a simples média dos respondentes havendo casos em que a opção não é muito clara. Sobre a aplicação do método, que acaba de descrever-se, a problemas de item de não resposta convem citar Lessler e Kalsbeek(1992): "*Although the weighting class adjustment method could conceivably be used to deal jointly with unit and*

item nonresponse, this option is seldom chosen since separate adjustments would be required for each data item."

Ajustamento por pós-estratificação

Enquanto no método anterior as células são constituídas nas primeiras fases da amostragem e dizem respeito a agrupamentos homogéneos de respondentes e de não respondentes, na pós-estratificação o agrupamento é feito depois de conhecidos os resultados da amostragem e a ideia é definir estratos, dentro dos quais os valores de Y sejam algo semelhantes, com base em variáveis de controle altamente corre- lacionadas com Y como, por exemplo, o número de pessoas por família. A dimensão dos estratos na população N_h pode ser conhecida antes do inquérito (por exemplo, através de um censo) mas a classificação dos elementos da amostra só pode fazer-se *a posteriori.*

Em resumo, o processo consiste em escolher uma (ou mais) variável de controle e tentar reajustar a sua distribuição na amostra por ponderação dos questionários preenchidos (veja-se Crespo(1993) nomeadamente no que respeita ao exemplo seguinte).

Exemplificando: pretende analisar-se uma amostra de 1000 agregados familiares e a estratificação é definida em termos do número de pessoas do

agregado. Recorrendo à informação censitária pode determinar-se para a amostra o efectivo teórico de cada estrato (veja-se Quadro 4.3).

Como pode ver-se pelo Quadro 4.3 o ajustamento da amostra é feito de modo que a distribuição dos questionários preenchidos por número de pessoas do agregado seja idêntica à da população. Os questionários a duplicar são obtidos tirando uma amostra ao acaso dos questionários preenchidos.

Quadro 4.3 - Exemplo de pós-estratificação

Agregados de	*Efectivo teórico*	*Questionários preenchidos*	*Questionários a duplicar*	*Questionários a destruir*
1 pessoa	182	99	83	-
2 pessoas	280	244	36	-
3 pessoas	218	190	28	-
4 pessoas	143	111	32	-
5 pessoas	90	80	10	-
6 pessoas ou mais	87	92	-	5
TOTAL	1000	816	189	5

Para evitar o enviesamento devido à não resposta é necessário que em cada um dos estratos, para a característica relevante, a média dos respondentes seja igual à média dos não respondentes, o que se consegue mais facilmente se a variável de controle (número de pessoas por agregado) for independente da atitude face ao inquérito.

Outras regras recomendam que os estratos sejam em número não muito elevado mas com razoável número de unidades (por exemplo, mais do que 20) em cada estrato.

Existem métodos mais complexos de ponderação (veja-se Lessler e Kalsbeek(1992)) mas o tema da pós-estratificação não vai desenvolver-se dada a semelhança que tem com a ponderação por classes.

Método de Deming

O método de Deming(1953) é utilizado para compensar a falta de resposta principalmente quando causada pela ausência do elemento seleccionado para a amostra. Consiste na realização de visitas sucessivas até que se consiga a resposta ou colaboração da unidade inicialmente não

respondente ou esta seja classificada como não respondente. Para aplicação do método a população é dividida em H estratos ou classes em função da probabilidade de resposta na primeira entrevista. Por exemplo, pode ter-se $H=3$ e as probabilidades de resposta na primeira entrevista serem respectivamente: *0,6*; *0,4 e 0,2*.

Seja:

ϖ_h - a proporção da população na classe h, $h=1,2,...,H$;

p_{ch} - a probabilidade de um respondente da classe h ser entrevistado até à c-ésima tentativa inclusivé, $c=1,2,...$ (antes ou durante o c-ésimo "*call-back*");

μ_h - a média da variável relevante na h-ésima classe da população;

σ_h^2 - a variância da variável relevante na h-ésima classe da população.

Por hipótese supõe-se que $p_{ch} > 0$ e que $E(\bar{y}_{ch}) = \mu_h$ onde $\bar{y}_{ch}$ é a média dos que responderam até à c-ésima tentativa inclusivé. Para a verdadeira média, $\bar{\mu}$, tem-se, evidentemente,

$$\bar{\mu} = \sum_{h=1}^{H} \varpi_h \mu_h \ .$$

Sendo n a dimensão da amostra, após a c-ésima tentativa os elemen-

tos da amostra podem arrumar-se em $H+1$ classes do seguinte modo:

- Classe $h=1$ - respondentes até à c-ésima tentativa - n_{c1}
- Classe $h=2$ - respondentes até à c-ésima tentativa - n_{c2}
- ...
- Classe $h=H$ - respondentes até à c-ésima tentativa - n_{cH}
- Subtotal ..... - n_c
- Classe $H+1$ - não respondentes até à c-ésima tentativa - $n_{c(H+1)}$
- Total..... - n

Se a correcção por se tratar de uma amostra de população finita[1] for ignorada, por um lado, o vector $(n_{c1}, n_{c2},..., n_{cH}, n_{c(H+1)})$ tem distribuição

Multinomial em n provas com probabilidades[2],

$$(\varpi_1 p_{c1}, \varpi_2 p_{c2}, ..., \varpi_H p_{cH}, 1-\sum_h \varpi_h p_{ch});$$

por outro lado, $n_c=n_{c1}+ n_{c2}+...+ n_{cH}$, número de entrevistas com "sucesso"

---

[1] Correcção que transforma a variância da média em populações infinitas por multiplicação pelo factor $(N-n)/N$.

[2] O somatório nas expressões abaixo é *entre* $h=1$ *e* $h=H$.

em n provas tem distribuição Binomial com probabilidade,

$$\sum_h \varpi_h p_{ch} .$$

Consequentemente,

(4.12) $$E(n_c) = n\sum_h \varpi_h p_{ch} ,$$

é o número esperado de entrevistas bem sucedidas até à c-ésima tentativa com uma amostra inicial de dimensão n.

Com n_c fixo, os elementos do vector $(n_{c1}, n_{c2}, ..., n_{cH})$ possuem também distribuição Multinomial com probabilidades,

$$(\varpi_1 p_{c1}/\sum_h \varpi_h p_{ch}, \varpi_2 p_{c2}/\sum_h \varpi_h p_{ch}, ..., \varpi_H p_{cH}/\sum_h \varpi_h p_{ch}) .$$

Consequentemente,

$$E(n_{ch}|n_c) = n_c \frac{\varpi_h p_{ch}}{\sum_h \varpi_h p_{ch}} ,$$

de modo que se $\bar{y}_c$ representa a média obtida com as observações recolhidas até à c-ésima tentativa, tem-se,

$$E(\bar{y}_c|n_c) = E(\sum_h n_{ch}\bar{y}_{ch}/n_c) ,$$

donde,

$$E(\bar{y}_c|n_c) = \frac{\sum_h n_c \varpi_h p_{ch} \mu_c}{n_c \sum_h \varpi_h p_{ch}}$$
$$= \frac{\sum_h \varpi_h p_{ch} \mu_c}{\sum_h \varpi_h p_{ch}}$$
$$= \bar{\mu}_c.$$

Como este resultado não depende de n_c tem-se para o valor esperado não condicionado $E(\bar{y}_c) = \bar{\mu}_c$. O enviesamento do estimador $\bar{y}_c$ é, portanto, $\bar{\mu}_c - \bar{\mu}$.

A variância condicional obtem-se por raciocínio semelhante,

(4.13) $$Var(\bar{y}_c|n_c) = \frac{\sum \varpi_c p_{ch} \left[\sigma_h^2 + (\mu_h - \bar{\mu}_c)^2\right]}{n_c \sum \varpi_h p_{ch}}.$$

Um valor aproximado (a menos de termos de ordem $1/n_c^2$) da variância não condicional obtem-se substituindo em (4.13) n_c pelo seu valor esperado dado por (4.12).

A solução final do problema envolve uma análise de custos (veja-se Cochran(1977) e Durão(1993)).

Métodos de imputação

Como foi diversas vezes referido, um dos maiores problemas com que se confrontam os investigadores que analisam os dados das sondagens é o de saber como tratar os dados que faltam ou os dados que se obtiveram e estão claramente errados.

Diversas estratégias podem ser seguidas para lidar com estes problemas, algumas das quais vão ser apresentadas. Tais estratégias são colectivamente referidas como **métodos de imputação**.

Em termos gerais são métodos que imputam, (isto é, preenchem) os valores dos itens que estão em falta. Podem considerar-se:

(a) métodos de **imputação simples**, em que cada valor em falta é substituído por um único valor imputado e a análise resultante trata o valor imputado como um verdadeiro valor observado. Este procedimento não tem em consideração uma certa incerteza em relação aos valores em falta;

(b) métodos de **imputação múltipla** cuja descrição se deixa a Rubin(1987): "*Multiple imputation is a statistical technique to take advantage of the flexibility in modern computing to handle missing data. With it, each missing value is replaced by two or more*

imputed values in order to represent the uncertainty about wich value to impute."

Levy e Lemeshow(1991) começam por referir dois métodos - apresentados a seguir - que "arrumam" em métodos de imputação, apesar de reconhecerem não o serem propriamente, uma vez que não preenchem explicitamente os valores em falta.

Apagar todos os registos que faltam ou elementos com dados incompletos

Este método consiste em apagar todas as unidades que não responderam a todos os itens. Uma vez que permite a utilização de métodos estatísticos para dados completos é um método bastante utilizado. Não é, porém, um bom método, pois pode diminuir consideravelmente a dimensão da amostra, que consequentemente perde qualidades estatísticas, além de que os estimadores surgem bastante enviesadas (pois naturalmente, como já foi destacado, os indivíduos que não respondem têm certamente um comportamento "estatístico" diferente dos que respondem).

Criação da categoria desconhecido

Com a criação desta categoria "desaparecem" os valores em falta e assim pode ser feito o estudo com métodos para dados completos. No

entanto, não se impede, tal como no caso anterior, a presença de enviesamentos, e pode abrir-se o caminho para uma interpretação confusa dos resultados.

Substituição da média

Em todas as unidades que não tenham respondido a determinado item, o respectivo valor é substituído pela média desse mesmo item obtida a partir de todas as unidades que responderam a esse mesmo item.

Ilustrando: Suponha-se, como no capítulo anterior, que de uma população de $N = 100$ unidades é retirada uma amostra de dimensão $n = 9$. A sondagem consiste na aplicação de um questionário composto por cinco itens pretendendo-se que as nove unidades amostrais respondam a todos os itens. Admita-se agora, por conveniência de exposição, que se obtém resposta das nove unidades, mas que algumas (três) não respondem a todos os itens (duas não respondem a um item e uma não responde a dois itens) conforme ilustrado na matriz seguinte, onde, como pode verificar-se, faltam os valores:

$$Y_{32},\ Y_{35},\ Y_{51} \quad e\ Y_{85}\ .$$

$$\begin{bmatrix} Y_{11} & Y_{12} & Y_{13} & Y_{14} & Y_{15} \\ Y_{21} & Y_{22} & Y_{23} & Y_{24} & Y_{25} \\ Y_{31} & - & Y_{33} & Y_{34} & - \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} & Y_{45} \\ - & Y_{52} & Y_{53} & Y_{54} & Y_{55} \\ Y_{61} & Y_{62} & Y_{63} & Y_{64} & Y_{65} \\ Y_{71} & Y_{72} & Y_{73} & Y_{74} & Y_{75} \\ Y_{81} & Y_{82} & Y_{83} & Y_{84} & - \\ Y_{91} & Y_{92} & Y_{93} & Y_{94} & Y_{95} \end{bmatrix}.$$

Decide-se então utilizar o método de imputação da média,

$$\hat{Y}_{51} = \frac{\sum_{i=1}^{4} Y_{i1} + \sum_{i=6}^{9} Y_{i1}}{8}, \quad \hat{Y}_{32} = \frac{\sum_{i=1}^{2} Y_{i2} + \sum_{i=4}^{9} Y_{i2}}{8}, \quad \hat{Y}_{35} = \hat{Y}_{85} = \frac{\sum_{i=1}^{2} Y_{i5} + \sum_{i=4}^{7} Y_{i5} + Y_{95}}{7},$$

como se mostra na matriz seguinte,

$$\begin{bmatrix} Y_{11} & Y_{12} & Y_{13} & Y_{14} & Y_{15} \\ Y_{21} & Y_{22} & Y_{23} & Y_{24} & Y_{25} \\ Y_{31} & \hat{Y}_{32} & Y_{33} & Y_{34} & \hat{Y}_{35} \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} & Y_{45} \\ \hat{Y}_{51} & Y_{52} & Y_{53} & Y_{54} & Y_{55} \\ Y_{61} & Y_{62} & Y_{63} & Y_{64} & Y_{65} \\ Y_{71} & Y_{72} & Y_{73} & Y_{74} & Y_{75} \\ Y_{81} & Y_{82} & Y_{83} & Y_{84} & \hat{Y}_{85} \\ Y_{91} & Y_{92} & Y_{93} & Y_{94} & Y_{95} \end{bmatrix}.$$

Como consequência, a estimativa da média da população para esse item ou variável é a mesma que se obtinha se tivessem sido consideradas apenas as unidades respondentes, coincidindo os resultados dessa estimativa com os obtidos no primeiro método apresentado.

Este método se, por um lado, tem a vantagem de imputar um valor esperado com uma certa credibilidade, tem, por outro lado, a desvantagem de fornecer variâncias e covariâncias ilusórias.

Métodos de imputação explícita

A referência a estes métodos é feita no plural dada a enorme diversidade de opções que se enquadram na imputação explícita embora tenham o objectivo comum de encontrar um "substituto" tão próximo quanto possível do "substituído".

O aspecto fundamental é o recurso a dados ou variáveis auxiliares e algumas das modalidades são: (1) quantificar uma escolha lógica para o valor "substituto" baseada nos dados auxiliares e na sua relação com a característica relevante Y; (2) estabelecer um modelo de regressão simples ou múltipla para predizer o item em falta a partir das variáveis auxiliares[1].

[1] Os modelos de regressão podem estudar-se de forma mais desenvolvida em Lessler e Kalsbeek(1992).

Segundo Kalton e Kasprzyk(1982) a imputação explícita pode descrever-se em termos gerais através de um modelo matemático,

(4.14) $$Z_i = f(X_{1i}, X_{2i}, ..., X_{ki}) + \xi_i \ ,$$

onde Z_i é o valor "substituto", $(X_{1i}, X_{2i}, ..., X_{ki})$ são as variáveis auxiliares (conhecidas), $f(.)$ uma função adequada e ξ_i uma variável residual a considerar eventualmente (se o valor substituto é encarado de um ponto de vista estocástico, em caso contrário ξ_i=0).

Um primeiro exemplo que, salvo melhor opinião, ilustra bem a imputação explícita pode encontrar-se nas Estatísticas de Cultura, Desporto e Recreio 1992 e 1993 publicadas pelo I.N.E..

A variável em causa diz respeito ao financiamento público das actividades culturais das Câmaras Municipais e é objecto de inquéritos por parte do I.N.E. desde 1986. O financiamento tem duas componentes: despesas correntes (remunerações e outras despesas) e despesas de capital.

Dado o carácter exaustivo do inquérito a falta de resposta atempada das Câmaras Municipais levanta necessariamente problemas.

Em 1992[1] o I.N.E resolveu a questão calculando os dados em falta para o ano de 1992 com base nos de 1991 de acordo com as seguintes taxas de correcção:

(a) despesas correntes: +11,8% (taxa de crescimento das despesas correntes da Administração Pública em 1992);

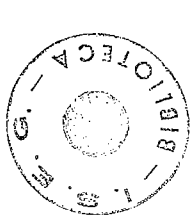
(b) despesas de capital: +8,9% (taxa de inflação em 1992).

Em 1993[2] a solução adoptada pelo I.N.E. foi a seguinte: (1) mantendo a distinção entre despesas correntes e despesas de capital, calculou índices de valor, tendo por base as Contas de Gerência das Câmaras Municipais em falta; (2) aplicou os índices de valor às respectivas rubricas do inquérito de 1992 admitindo, portanto, que os gastos globais e os gastos em cultura evoluem da mesma forma (em valor); (3) na hipótese de ser impossível o cálculo de índices de valor como indicado no ponto (1), as despesas correntes são estimadas recorrendo ao Índice de Valor do Consumo Colectivo das Administrações Públicas e as despesas de capital recorrendo ao Índice de Valor do Investimento da Administração Local (valores do Orçamento Geral do Estado 93/94, estimativa de execução).

---

[1] Câmaras Municipais estimadas: Castelo Branco, Lisboa, Rio Maior, Palmela, Seixal, Setúbal, Olhão e Porto Santo.

[2] Câmaras Municipais estimadas: Almeirim, Bombarral, Cartaxo, Madalena, Rio Maior, Sesimbra, Sintra, Setúbal, Santa Cruz e São Vicente.



Um segundo exemplo, menos interessante, consta das Estatísticas do Ambiente também publicadas pelo I.N.E. Os dados físicos (utilização dos solos, sistemas de abastecimento de água, drenagem e tratamento de águas residuais, etc.) referentes a 1991 tiveram de ser publicados em 1994, atraso que ficou a dever-se ao esforço feito no sentido de realizar uma recolha anual com uma taxa de resposta de 100%, por forma a construir uma base de cálculo para suprir a falta de resposta em futuros levantamentos. Mesmo com tal atraso o I.N.E. não conseguiu dados com a qualidade desejada. Em 1993 notaram-se algumas melhorias mas não foi possível eliminar a falta de resposta cuja taxa ultrapassou os 20% no que diz respeito aos Açores e Madeira. No entanto, para dispor de dados globais, às Câmaras Municipais em falta foram imputados os valores referentes a 1991, apenas tendo sido actualizados os dados relativos à população residente. Como pode ler-se nas Estatísticas do Ambiente para 1993: "*A melhoria da qualidade dos dados da edição de 1993 relativamente à anterior, permitiu salientar, e consequentemente corrigir após confirmação junto das câmaras municipais, algumas inconsistências verificadas na edição de 1991*".

Quando a função $f(.)$ é linear tem-se como caso particular de (4.14),

$$Z_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_{ki} X_{ki} + \xi_i \ , \tag{4.15}$$

que pode tratar-se, pelos métodos habituais, como uma equação de regressão linear múltipla[1].

As potencialidades da regressão linear são bem conhecidas e não vão alargar-se as considerações em torno das suas aplicações ao caso da falta de resposta. Um modalidade interessante de imputação explícita foi recentemente sugerida por Efron(1994) num caso em que se recolheram os resultados dos exames de 22 estudantes nas disciplinas A, B, C, D e E e em que se perderam algumas notas nas disciplinas A e E. Para preencher os valores em falta Efron utiliza um modelo linear "*two-way*",

$$Y_{ij} = \mu + \alpha_i + \beta_j \ ,$$

com $i=1,2,...,22$ (esudantes) e $j=1,2,3,4,5$ (disciplinas) que, evidentemente, ajusta apenas às células observadas, minimizando,

$$\sum (Y_{ij} - \mu - \alpha_i - \beta_j)^2 ,$$

com as condições,

$$\sum_{i=1}^{22} \alpha_i = 0 \quad , \quad \sum_{j=1}^{5} \beta_j = 0 ,$$

---

[1] Sugere-se também a consulta do "*método de Buck*" abreviado em Särndal Wretman e Swesson(1992) e mais desenvolvido em Rubin e Little(1987)

de forma a estimar primeiro $\hat{\mu}$, $\hat{\alpha}_i$ e $\hat{\beta}_j$ e de seguida os valores em falta,

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j \ .$$

A imputação da média pode ser também encarada como um caso específico da imputação por regressão onde as variáveis preditor são variáveis *dummy* indicadoras para as células dentro das quais a média é imputada.

Para tratar a imputação por regressão de um ponto de vista mais genérico - seguindo parcialmente Rubin e Little(1987) - vai adoptar-se a abordagem Bayesiana com as "suas" distribuições *a priori* e *a posteriori*. Em primeiro lugar deve atender-se à notação utilizada. Representa-se o conjunto de valores incluídos na amostra por Y_{inc}, o qual se subdivide em dois conjuntos - Y_{obs} e Y_{mis} - que designam respectivamente os valores que pertencendo à amostra são observados e os valores que pertencendo à amostra não são observados (estão em falta); enfim designam-se por X_i, $(i=1,2,...,n)$ as variáveis auxiliares que são registadas (observadas) para todas as unidades amostrais.

Observada uma amostra aleatória simples de dimensão n, $(Y_i, X_{i1},...,X_{ik})$, $i=1,2,...,$em que a incidência de resposta não depende de Y, ser condicionado pelos X_i $(i=1,2,...,n)$, Rubin e Little(1987) assumem que,

a)

$$(Y_i|X_{i1},...,X_{ik},\theta) \sim N(\mu_i,\sigma^2\upsilon_i),$$

$$\mu_i = \beta_0 + \sum_j \beta_j \mu_{ij},$$

onde, $\upsilon_i = \upsilon(X_{i1},...,X_{ik})$ é uma função conhecida que caracteriza a heterogeneidade da variância;

b)

$$\theta = (\hat{\beta}_0,...,\hat{\beta}_k,\ln\sigma^2),$$

tem distribuição *a priori* localmente uniforme.

Os valores das unidades não respondentes são estimados pela sua média *a posteriori*,

$$E(Y_i|Y_{obs},X_{inc}) = \hat{\beta}_0 + \sum_j \hat{\beta}_j \mu_{ij},$$

onde $(\hat{\beta}_0,...\hat{\beta}_k)$, são coeficientes obtidos pelo método dos mínimos quadrados ponderados, ponderando a i-ésima unidade por υ_i^{-1}. Para desenvolver este modelo podem consultar-se Rubin e Little(1987) e Royall and Herson(1983), entre outros.

Imputação por "*Hot deck*"

As faltas de resposta são substituídas por valores seleccionados de respondentes com comportamentos semelhantes na sondagem corrente, evitando o elevado grau de subestimação da variância inerente aos métodos

de imputação que preenchem valores com grandes médias ou médias de subdomínios.

As unidades são agrupadas em células que são subconjuntos da amostra (ou da população) formados de forma exclusiva, exaustiva e tão homogéneos quanto possível, com base numa característica comum que pode resultar de uma classificação cruzada de atributos demográficos ou outros. Dentro das células as unidades devem ser convenientemente organizadas; como dizem Levy e Lemeshow(1991): "*In the hot deck method, the original data file is sorted in such a way that the order of the individual records corresponds to the structure of the sample design (e.g., individuals from the same cluster might be together in the file)*". Por exemplo, se a sondagem que se está a efectuar é aplicada a domésticas apenas, as células podem ser separadas pelo número do bairro fiscal, em que a primeira célula corresponde ao primeiro bairro, a segunda ao segundo e assim sucessivamente.

Para cada célula organiza-se um **registo** contendo os valores de um indivíduo que respondeu a todos os itens. Ao fazer uma passagem pela lista das unidades, a célula a que cada indivíduo pertence é identificada e, das duas, uma: (1) se o indivíduo não respondeu a determinada questão ou não indicou o valor da variável em causa, os dados em falta são substituídos pelos dados do **registo;** (2) se o indivíduo respondeu a todos os itens, então

os valores com ele associados passam a ocupar o **registo**, em substituição dos valores do indivíduo "completo" anterior.

Considere-se a situação usada a propósito da imputação da média para ilustrar o método do "*hot deck*". Suponha-se que se formam duas células sendo a primeira coluna que as distingue. A primeira célula é constituída pelas cinco primeiras linhas e a segunda pelas quatro restantes,

isto é,

$$Y_{11} = Y_{21} = Y_{31} = Y_{41} = Y_{51} = Y_A \quad , \quad Y_{61} = Y_{71} = Y_{81} = Y_{91} = Y_B \ .$$

Tem-se, então,

$$\begin{bmatrix} Y_A & Y_{12} & Y_{13} & Y_{14} & Y_{15} \\ Y_A & Y_{22} & Y_{23} & Y_{24} & Y_{25} \\ Y_A & - & Y_{33} & Y_{34} & - \\ Y_A & Y_{42} & Y_{43} & Y_{44} & Y_{45} \\ Y_A & - & Y_{53} & Y_{54} & Y_{55} \\ Y_B & Y_{62} & Y_{63} & Y_{64} & Y_{65} \\ Y_B & Y_{72} & Y_{73} & Y_{74} & Y_{75} \\ Y_B & Y_{82} & Y_{83} & Y_{84} & - \\ Y_B & Y_{92} & Y_{93} & Y_{94} & Y_{95} \end{bmatrix} .$$

O lugar "vago" que deveria ser ocupado por Y_{32} é ocupado por Y_{22} ; o que deveria ser ocupado por Y_{35} é ocupado por Y_{25}; o que deveria ser ocupado

por Y_{52} é ocupado por Y_{42} e, finalmente, o que deveria ser ocupado por Y_{85} é ocupado por Y_{75} . Obtém-se, consequentemente uma matriz de dados completa,

$$\begin{bmatrix} Y_A & Y_{12} & Y_{13} & Y_{14} & Y_{15} \\ Y_A & Y_{22} & Y_{23} & Y_{24} & Y_{25} \\ Y_A & Y_{22} & Y_{33} & Y_{34} & Y_{25} \\ Y_A & Y_{42} & Y_{43} & Y_{44} & Y_{45} \\ Y_A & Y_{42} & Y_{53} & Y_{54} & Y_{55} \\ Y_B & Y_{62} & Y_{63} & Y_{64} & Y_{65} \\ Y_B & Y_{72} & Y_{73} & Y_{74} & Y_{75} \\ Y_B & Y_{82} & Y_{83} & Y_{84} & Y_{75} \\ Y_B & Y_{92} & Y_{93} & Y_{94} & Y_{95} \end{bmatrix},$$

podendo então aplicar-se os procedimentos usuais para dados completos.

Existem diversas variantes deste método e um vasto número de procedimentos "*Hot deck*" têm sido propostos, incluindo *random overall imputation, random imputation within classes, sequencial hot deck imputation, hierarchial hot deck imputation, distance function matching* . Uma breve referência pode ler-se em Särndal, Swensson e Wretman(1992).

Imputação e "Hot deck" nos XII e XIII Recenseamentos Gerais da População[1]

A falta de resposta também se revela altamente inconveniente nos censos cujo objectivo é um levantamento exaustivo de todas as unidades e não um levantamento "truncado" que torne difícil o apuramentro dos dados globais. Em geral a imputação dos valores em falta é feita conjuntamente com uma operação de validação. Como pode ler-se nos "Antecedentes, Metodologia e Conceitos" o I.N.E., na fase de **correcção automática**[2], criou um sistema mecânico complexo de correcção, variável a variável, actuando quando se verificava alguma incoerência que se decidiu não corrigir manualmente para não sobrecarregar este trabalho ou quando se detectou a **ausência de valor,** normalmente por ausência de resposta ao questionário. A correcção foi feita pelo computador no preciso momento em que se identificou qualquer daquelas deficiências.

Duas premissas constituiram, por assim dizer, a segurança do sistema:

---

[1] A presente secção baseia-se no volume "Antecedentes, Metodologia e Conceitos", do XII Recenseamento Geral da População, II Recenseamento Geral da Habitação, 1981, publicado pelo I.N.E. em 1984 e nos elementos gentilmente fornecidos pelo Dr.Fernando Casimiro e Dra.Helena Gomes.

[2] Esta fase, no esquema geral de tratamento da informação, foi antecedida pela análise e codificação de questionários, pelo registo em suporte informático, pela validação automática e pela correcção manual.

(a) a existência, como base de apoio, de um conjunto de variáveis fundamentais cujos valores foram sujeitos com antecedência ao processo de validação e correcção manual;

(b) a frequência de correcção de cada variável não era elevada, isto é, pretendeu-se não atingir 10% de valores corrigidos em cada variável.

As variáveis fundamentais foram:

No edifício: número de alojamentos;

No alojamento: tipo, número de ocupantes, regime de propriedade/aluguer ou ocupação.

No indivíduo: sexo, idade, nacionalidade e situação perante a residência.

A partir destas variáveis corrigiram-se as restantes, uma a uma segundo uma sequência definida de modo a tirar o máximo partido das combinações adequadas das variáveis que iam ficando corrigidas. A solução normal é a de imputar sempre um valor válido e considerado o mais correcto no conteúdo da restante informação, recorrendo ao **método do "*hot deck*".**

O método do "*hot deck*" foi aplicado através da construção de matrizes, com duas ou três entradas, colocando nas colunas os valores de uma variável auxiliar e nas linhas os valores da outra ou outras variáveis auxiliares. Os elementos da matriz são os diversos valores "impostos" à variável a corrigir ou preencher em função das variáveis fundamentais consideradas. Cada matriz foi preenchida incialmente com os valores lógicos para cada cruzamento das variáveis auxiliares.

Quadro 4.4 - Extracto da matriz inicial do método "*hot deck*"

(correcção ou preenchimento do "número de pavimentos")

Nº de Alo-jamentos	*- 1919*	*Época de Construção 1919/45*	*...*	*1976/79*	*1980/81*
1	1	1	...	1	1
2	2	2	...	2	2
3	2	2	...	2	2
...	...	...	...	...	...
20/29	6	6	...	7	7
30/39	8	8	...	9	9
40+	10	10	...	10	10

No Quadro 4.4 apresenta-se um extracto da matriz com os valores de iniciais para correcção da variável "número de pavimentos" em função da época da construção do edifício e do número de alojamentos.

Ao analisar cada variável, se esta tem um valor válido, este vai alimentar o "*hot deck*", isto é, substitui o valor existente na casa da matriz que lhe corresponde. Pelo contrário, se a variável não tem qualquer valor ou tem um valor inaceitável, não só não há lugar a qualquer alteração da matriz, como se imputará à variável em análise o valor do elemento da matriz correspondente à intersecção dos valores das variáveis auxiliares que caracterizam a unidade em questão.

Convém reparar que o método de imputação assenta num processo perfeitamente aleatório, em que a probabilidade de um valor a imputar está directamente relacionada com a sua frequência de entrada na matriz. Quer dizer, para cada variável, a distribuição dos valores corrigidos é mesma que se observa nos valores válidos dessa mesma variável. Dada a pequena percentagem de valores a corrigir, no Censo de 1981 verificou-se uma grande alimentação do "*hot deck*" (substituição extremamente frequente dos valores da matriz) o que deu uma importância muito reduzida à matriz de inicialização.

Para ilustar, em termos esquemáticos, o funcionamento do "*hot deck*" no Censos 81 considere-se a questão 7 do Questionário Individual cuja estrutura se indica no Quadro 4.5.

No Questionário Individual encontra-se a Questão 6 (Data de Nascimento) que permite determinar a idade do indivíduo e no Questionário de Família a Questão B a preencher com a Lista da Família a Residir no Alojamento e contendo eventualmente o nome do pai e/ou da mãe e o nome do conjuge e, para conferência, pede-se a indicação do respectivo nº de ordem (NO) em colunas separadas.

Quadro 4.5 - (7) Estado Civil

Estado Civil:	
Solteiro (a)	1
Casado (a)	2
Viuvo (a)...	3
Separado (a)	4
Divorciado (a)	5

No presente exemplo o "*hot deck*" funciona do seguinte modo:

1) O indivíduo tem idade inferior ou igual a 14 anos:

1.1) declara ser solteiro $\Rightarrow$ vai alimentar ***HD*;**

1.2) declara outro estado civil ⇐ faz-se a correcção para solteiro.

2) O indivíduo tem idade superior a 14 anos e indicou NO do conjuge:

2.1) declara ser casado ⇒ vai alimentar ***HD*;**

2.2) declara outro estado civil ⇐ faz-se a correcção para casado.

3) O indivíduo tem idade superior a 14 anos, não indicou NO do conjuge, mas indicou NO do pai e/ou mãe:

3.1) declara ser solteiro ⇒ vai alimentar ***HD*;**

3.2) declara outro estado civil ⇐ faz-se a correcção para solteiro:

4) O indivíduo tem idade superior a 14 anos, não indicou NO do conjuge, nem indicou NO do pai e/ou da mãe:

4.1) declara um qualquer estado civil ⇒ vai alimentar ***HD;***

4.2) não responde ⇒ o estado civil é prenchido por ***HD.***

No Censos 91 houve muitas melhorias[1] que não podem aqui abordar-se. A validação foi feita a cinco níveis sendo no nível III que se recorreu ao método do "*hot deck*". Assim, a nível III, o I.N.E. procedeu ao controlo e validação dos dados já registados e validados a nível I e II, para tratar

[1] Por exemplo, em relação aos recenseamentos anteriores verificou-se uma profunda modificação no tratamento da informação das respostas e perguntas abertas (designações de âmbito geográfico, de cursos, de profissões e ramos de actividade)

algumas ausências de respostas, assim como detectar incoerência de informação por respostas deficientes às perguntas em causa ou erro de digitação. Em particular, foram validadas as variáveis relacionadas com o sexo, a residência habitual, a data de nascimento, as relações de parentesco e os números de ordem do conjugue, do pai e da mãe. Estas variáveis foram consideradas fundamentais e constituiram a base das validações dos níveis seguintes.

O nível III compõe-se de três tipos de correcções:

- verificações de coerência por via manual (consulta de questionários) admitindo que o erro possa estar a montante;
- determinística: imputação automática de valores quando a aceitação prévia de determinados pressupostos implica uma única resposta pré-definida ;
- probabilística: correcções automáticas com base no recurso ao "*hot deck*" para atribuição de valores no caso de ausência de resposta, incoerência ou incorrecção. O método do "*hot deck*" é concebido, mais uma vez, em duas etapas: na primeira, preparam-se as matrizes de valores pré-definidos com base em variáveis consideradas pertinentes para a definição dos valores a atribuir; na segunda, a matriz vai sendo alimentada com os valores válidos a que se recorre nos casos de incoerência ou falta de resposta.

As validações automáticas do nível IV e V foram feitas no computador central.

Ao nível IV procedeu-se à verificação das incoerências com correcções automáticas determinísticas, isto é, às variáveis que não estavam preenchidas ou que apresentavam incoerências imputaram-se valores lógicos derivados das respostas já existentes.

A nível V procedeu-se à verificação de incoerências ou falta de respostas com correcções probabilísticas do tipo já indicado.

Para ilustar, em termos esquemáticos, o funcionamento do "*hot deck*" no Censos 91 considere-se a questão A12 do Questionário de Alojamento cuja estrutura se indica no Quadro 4.6.

Quadro 4.6 - Sistema de esgotos

- O alojamento tem sistema de esgotos:
  - Ligado à rede pública..... 2
  - Sistema particular (fossa séptica, etc.)..... 4
  - Outros (fossa aberta, vala)..... 6
- O alojamento não tem sistema de esgotos..... 8

A montante de A12 havia ainda duas questões:

- Existem indivíduos no alojamento:

  Sim

  Não

- A subsecção[1] tem rede de esgotos:

  Sim

  Não

- Se a subsecção tem rede de esgotos, a resposta à questão A12 pode assumir qualquer valor (2, 4, 6, 8). Se houver falta de resposta procura-se o valor pelo sistema "*hot deck*" no ***HD1A12*** ⇒ para subsecções que tenham rede pública.

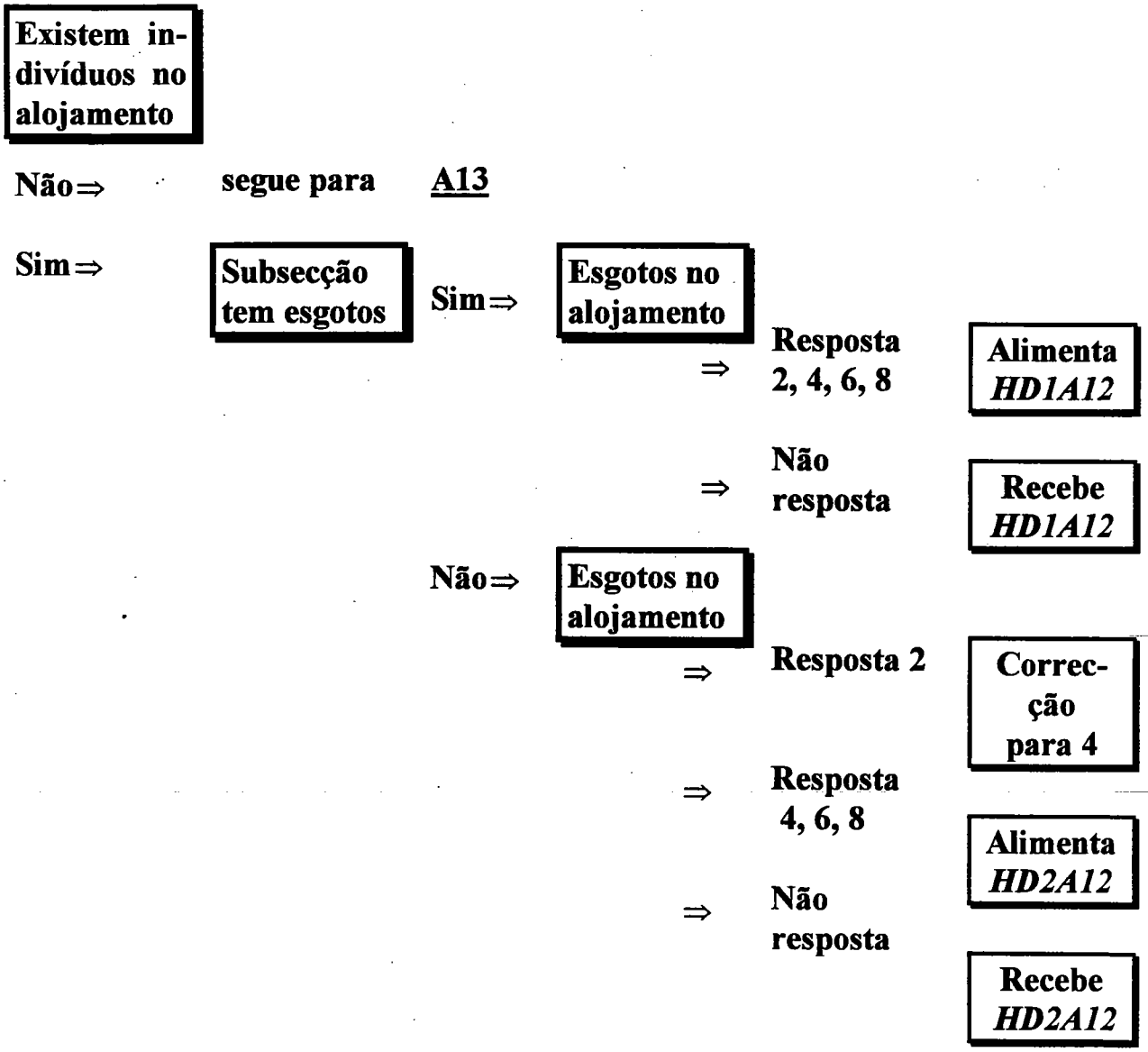
- Se a subsecção não tem rede de esgotos não se admite que o alo-jamento seja servido por rede pública. Se é essa a resposta (2) faz-se a correcção substituindo por "sistema particular" (4). Se houver falta de resposta procura-se o valor pelo sistema "*hot deck*" no ***HD2A12*** ⇒ para subsecções que não tenham rede pública.

---

[1] As freguesias são divididas em secções (300 alojamentos) e em subsecções (quarteirões nos meios urbanos, lugares nos meios rurais).

Quadro 4.7 - Esquema da aplicação do "hot deck"

na questão A12 - Sistema de esgotos



• As respostas coerentes vão alimentar, no primeiro caso, ***HD1A12***, no segundo caso, ***HD2A12***.

Para dar uma ideia mais completa pode ver-se o o esquema geral do Quadro 4.7.

Método de imputação "*Cold deck*"

Este método é, no fundo, uma variante do anterior, pois também faz a identificação de células. No entanto, hoje em dia, é raramente usado na prática, porquanto utiliza imputações baseadas noutras fontes que não as sondagens correntes, por exemplo, sondagens anteriores semelhantes feitas à mesma população.

Ilustrando: considere-se mais uma vez o caso que tem servido sistematicamente para exemplificação e suponha-se que são os resultados de uma sondagem efectuada num determinado ano e ainda que no ano anterior a mesma sondagem tinha sido feita às mesmas empresas tendo-se obtido a matriz completa $Y^* = (Y_{ij}^*)$, $i=1,...,9$; $j=1,...,5$. Assim a matriz dos dados completos como resultado da utilização do método de imputação

por "Cold deck" é a seguinte:

$$\begin{bmatrix} Y_{11} & Y_{12} & Y_{13} & Y_{14} & Y_{15} \\ Y_{21} & Y_{22} & Y_{23} & Y_{24} & Y_{25} \\ Y_{31} & Y_{32}^{*} & Y_{33} & Y_{34} & Y_{35}^{*} \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} & Y_{45} \\ Y_{51} & Y_{52}^{*} & Y_{53} & Y_{54} & Y_{55} \\ Y_{61} & Y_{62} & Y_{63} & Y_{64} & Y_{65} \\ Y_{71} & Y_{72} & Y_{73} & Y_{74} & Y_{75} \\ Y_{81} & Y_{82} & Y_{83} & Y_{84} & Y_{85}^{*} \\ Y_{91} & Y_{92} & Y_{93} & Y_{94} & Y_{95} \end{bmatrix}.$$

Quando há vários valores que podem servir para a imputação, um é escolhido aleatoriamente.

Imputação múltipla

A imputação múltipla refere-se ao procedimento de substituir cada valor em falta ou cada valor deficiente (valor cuja validade possa ser posta em causa por constar, por exemplo, da resposta a um questionário, haver uma pessoa com trinta anos de idade e com um filho de sessenta anos) por um ou mais valores aceitáveis, representando uma distribuição de possibilidades. Isto é, para cada "casa" vaga é proposto um vector de $m \geq 2$ valores imputáveis, em que os m valores são ordenados no sentido de se poder criar m conjuntos completos de dados. Assim, os valores a imputar formam uma matriz $k \times m$ onde k é o total dos itens em falta. A primeira

coluna, quando substitui os dados em falta, forma o primeiro conjunto de dados completos; a segunda coluna, quando substitui os dados em falta , dá origem a um segundo conjunto de dados completos e assim por diante.

Ilustrando com o exemplo que tem vindo a acompanhar a exposição, considere-se um caso de imputação múltipla em que sucessivamente se imputam a cada valor em falta dois valores ($m=2$), isto é constroi-se um vector com dois valores a imputar para cada valor em falta. Assim, para o elemento Y_{13} em falta constroi-se o vector de imputações $[Z_{11} \quad Z_{12}]$, procedendo-se do mesmo modo com Y_{32}, Y_{34} e Y_{85}.

Os valores a imputar são guardados numa matriz auxiliar com uma linha para cada valor em falta:

$$\begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \\ Z_{31} & Z_{32} \\ Z_{41} & Z_{42} \end{bmatrix},$$

Imputando a primeira coluna desta matriz aos valores em falta obtém-se

uma matriz completa:

$$\begin{bmatrix} Y_{11} & Y_{12} & Y_{13} & Y_{14} & Y_{15} \\ Y_{21} & Y_{22} & Y_{23} & Y_{24} & Y_{25} \\ Y_{31} & Z_{11}\hookleftarrow & Y_{33} & Y_{34} & Z_{12}\hookleftarrow \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} & Y_{45} \\ Y_{51} & Z_{13}\hookleftarrow & Y_{53} & Y_{54} & Y_{55} \\ Y_{61} & Y_{62} & Y_{63} & Y_{64} & Y_{65} \\ Y_{71} & Y_{72} & Y_{73} & Y_{74} & Y_{75} \\ Y_{81} & Y_{82} & Y_{83} & Y_{84} & Z_{14}\hookleftarrow \\ Y_{91} & Y_{92} & Y_{93} & Y_{94} & Y_{95} \end{bmatrix}$$

Imputando a segunda coluna da matriz auxiliar obtém-se a segunda matriz completa:

$$\begin{bmatrix} Y_{11} & Y_{12} & Y_{13} & Y_{14} & Y_{15} \\ Y_{21} & Y_{22} & Y_{23} & Y_{24} & Y_{25} \\ Y_{31} & Z_{21}\hookleftarrow & Y_{33} & Y_{34} & Z_{22}\hookleftarrow \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} & Y_{45} \\ Y_{51} & Z_{23}\hookleftarrow & Y_{53} & Y_{54} & Y_{55} \\ Y_{61} & Y_{62} & Y_{63} & Y_{64} & Y_{65} \\ Y_{71} & Y_{72} & Y_{73} & Y_{74} & Y_{75} \\ Y_{81} & Y_{82} & Y_{83} & Y_{84} & Z_{24}\hookleftarrow \\ Y_{91} & Y_{92} & Y_{93} & Y_{94} & Y_{95} \end{bmatrix}$$

É claro que não deve construir-se uma matriz auxiliar de imputações com maior dimensão do que a matriz original dos dados. Segundo diversos autores é conveniente ter-se $(2 \leq m \leq 10)$, aliás pode ler-se em Lessler e

Kalsbeek(1992) que: "*A simulation study by Rubin and Schenker(1986) suggests, for example, that two sets of imputed values may be sufficient for variables with moderate rates of item nonresponse and that three sets would be satisfactory for larger item nonresponse rates.*"

Cada conjunto de dados completos é analisado através dos procedimentos clássicos para dados completos, como se os dados imputados fossem os dados reais obtidos dos não respondentes; este tipo de análise (de um modo geral) ignora a distinção entre respondentes e não respondentes. Obtêm-se deste modo, de cada conjunto de dados, estimativas para os parâmetros em estudo. Combinando o resultado das inferências consegue-se que estas traduzam a incerteza acerca dos valores em falta. Quando os m conjuntos de imputações são esquemas aleatórios repetidos sob um modelo, as m inferências de dados completos podem ser combinadas para formar uma inferência que de um modo próprio reflecte a incerteza devida à falta de resposta sob aquele modelo.

A imputação múltipla de dados pode ser utilizada para simular a distribuição do estimador sob um método particular, assumindo que o método envolve uma espécie de componentes aleatórias. Assim, os valores para um não respondente irão diferir. A dificuldade prática deste método é a necessidade de produzir e reter mais do que um valor "dador" para cada valor em falta.

Quando as imputações são de dois ou mais modelos de não resposta, as inferências combinadas sob os modelos podem ser contrastadas através de métodos adequados.

A capacidade para usar métodos de análise para dados completos e a de incorporar o conhecimento daquele que recolhe os dados - ao seleccionar os valores a imputar - são duas vantagens comuns à imputação simples e à imputação múltipla em relação aos restantes métodos.

Segundo Rubin(1987), a imputação múltipla tem em relação à simples as vantagens de: (1) aumentar a eficiência de estimação quando as imputações são esquematizadas aleatoriamente numa tentativa de representar a distribuição dos dados; (2) apresentar inferências que reflectem a variabilidade adicional devido aos valores em falta sob o modelo para a não resposta quando a imputação múltipla representa repetidos esquemas aleatórios e, ainda nessa situação; (3) permitir a estreiteza do estudo da sensibilidade das inferências a vários modelos de não resposta, simplesmente utilizando repetidamente métodos para dados completos.

Por outro lado existem três desvantagens óbvias da imputação múltipla em relação à imputação simples, pois a imputação múltipla exige mais trababalho do que a imputação simples, assim como é mais trabalhoso analisar um conjuto de dados multi-imputados do que dados com

imputação simples. É também necessário mais espaço em computador para armazenar um conjunto de dados multi-imputados.

O exemplo que se apresenta a seguir, adaptado de Rubin(1987), é particularmente elucidativo. Considere-se uma amostra de $n=10$ unidades recolhida de uma população com $N=1000$ unidades. O objectivo é estimar a partir da amostra a média da população Y=dimensão em 1990, usando como variável auxiliar X=dimensão em 1980. Duas unidades recusam-se a responder. Parte-se do princípio de que se a amostra só tivesse respondentes empregava-se como estimador o rácio $\overline{X}(\overline{y}/\overline{x})$ onde $\overline{X}$ é a média dos valores de X na população, $\overline{y}$ e $\overline{x}$ as médias de Y e X respectivamente na amostra. Considerando a expressão,

$$\hat{v} = \frac{\sum \left[Y_i - (X_i \overline{y}/\overline{x})\right]^2}{n-1} ,$$

podia então construir-se para $\overline{Y}$ um intervalo de confiança (aproximado) a 95% (por exemplo),

$$\overline{X}(\overline{y}/\overline{x}) \pm 1{,}96\ (\hat{v}/n)^{1/2} .$$

Os dados (artificiais) encontram-se no Quadro 4.8 No Quadro 4.9 encontram-se quatro imputações para cada um dos dois casos de falta de resposta.

Quadro 4.8 - Dados observados

Unidades	Y_i	X_i	Unidades	Y_i	X_i
1	10	8	6	15	18
2	?	9	7	20	6
3	14	11	8	4	4
4	?	13	9	18	20
5	15	16	10	22	25

Quadro 4.9 - Imputação múltipla

Modelo	**I**		**II**	
Repetição	**1**	**2**	**1**	**2**
Unidade 2	10	14	12	17
Unidade 4	16	14	19	17

As imputações múltiplas foram obtidas com dois modelos[1], com duas repetições saídas da aplicação de um método muito semelhante ao "*hot deck*" (veja-se Rubin(1987)). Os quatro conjuntos de dados completos constam dos Quadros 4.10-4.13.

[1] O modelo (I) parte da hipótese de que um respondente e o não respondente comportam-se do mesmo modo caso tenham o mesmo valor de X; o modelo (II) considera um comportamente diferenciado entre respondentes e não respondentes mesmo que tenham o mesmo valor de X.

Quadro 4.10 - Dados observados (modelo I, rep.1)

Unidades	Y_i	X_i	Unidades	Y_i	X_i
1	10	8	6	15	18
2	10	9	7	20	6
3	14	11	8	4	4
4	16	13	9	18	20
5	15	16	10	22	25

Quadro 4.11 - Dados observados (modelo I, rep.2)

Unidades	Y_i	X_i	Unidades	Y_i	X_i
1	10	8	6	15	18
2	14	9	7	20	6
3	14	11	8	4	4
4	14	13	9	18	20
5	15	16	10	22	25

Quadro 4.12 - Dados observados (modelo II, rep.1)

Unidades	Y_i	X_i	Unidades	Y_i	X_i
1	10	8	6	15	18
2	12	9	7	20	6
3	14	11	8	4	4
4	19	13	9	18	20
5	15	16	10	22	25

Quadro 4.13 - Dados observados (modelo II, rep.2)

Unidades	Y_i	X_i	Unidades	Y_i	X_i
1	10	8	6	15	18
2	17	9	7	20	6
3	14	11	8	4	4
4	17	13	9	18	20
5	15	16	10	22	25

A análise processa-se evidentemente com métodos adequados ao tratamento de dados completos. No Quadro 4.14 encontram-se as estimativas e as variâncias associadas com cada conjunto de dados completos. Os estimadores da média são $\overline{X}(\overline{y}/\overline{x})$ e as respectivas variâncias são $\hat{v}/n$. Seguidamente combinam-se os resultados obtidos de acordo com cada um dos modelos em termos de inferências sobre $\overline{Y}$ conforme se mostra no Quadro 4.15. Finalmente, a seguir ao Quadro 4.12 apresentam-se os intervalos de confiança para $\overline{Y}$ cujo centro é a média das estimativas , seja,

$$13.48=(13.38+13.57)/2 \quad \text{e} \quad 13.99=(13.85+14.12)/2 \ ,$$

e cuja amplitude é calculada considerando a variância nas suas duas componentes:

(a) a média da variância intra-imputação, seja,

$$(2.96+3.19)/2 \quad \text{e} \quad (3.38+3.84)/2,$$

(b) a variância entre-imputação, seja,

$\{(13.38-13.48)^2+(13.57-13.48)^2\}$ e $\{(13.85-13.99)^2+(14.12-13.99)^2\}$

e aplicando a fórmula,

estimativa total da variância=média da variância intra-imputação+

$+(1+m^{-1})$ *× estimativa da variância entre-imputação,*

onde o factor $(1+m^{-1})$ é um ajustamento justificado pelo emprego de um número finito ($m=2$) de imputações.

Quadro 4.14- Estimativas da média e da variância associada

Modelo	**I**		**II**	
Repetição	**1**	**2**	**1**	**2**
Média	13.38	3.57	3.85	4.12
Variância	2.96	3.19	3.38	3.84

Quadro 4.15 - Estimativas combinadas

	Modelo I	**Modelo II**
Média	13.48	13.99
Variância	3.10	3.66

- Intervalo de confiança a 95% para $\overline{Y}$ decorrente do modelo I:

(10.0, 16.9);

- Intervalo de confiança a 95% para $\overline{Y}$ decorrente do modelo II:

(10.2, 17.7).

Para terminar vão considerar-se com Y_i univariada[1] alguns exemplos de procedimentos de imputação múltipla começando por considerar o modelo de regressão normal com Y_i

$$Y_i \sim N(X_i\beta, \sigma^2),$$

que consiste na especificação de,

$$f(Y_i|X_i,\theta); \quad \theta = (\beta, \log\sigma),$$

com, β vector e σ escalar.

A tarefa de modelização é completada intoduzindo a habitual distibuição *a priori* imprópria para θ, isto é,

$$h(\theta) \propto constante$$

e supondo que $n_1 > q$.

Ao abordar a estimação, deve ter-se presente que a distribuição *a posteriori* de θ envolve a matriz X e apenas os valores Y_i observados, pois,

$$h(\theta|X, Y_{obs}) = \frac{\left[\prod_{obs} f_{Y|X}(Y_i|X_i,\theta)\right]h(\theta)}{\int\left[\prod_{obs} f_{Y|X}(Y_i|X_i,\theta)\right]h(\theta)d\theta}.$$

Por um lado (veja-se Box e Tiao(1973)) sabe-se que *a posteriori* σ^2

---

[1] Y_i univariada com variáveis auxiliares é uma situação bastante mais geral do que aparenta ser, pois muitos casos de imputação múltipla com Y_i multivariada podem criar-se por aplicação repetida de métodos estruturados para Y_i univariada.

tem distribuição,

(4.16) $$\frac{\hat{\sigma}_1^2(n_1-q)}{g},$$

onde g é uma variável aleatória com distribuição de $\chi^2_{n_1-q}$.

Por outro lado, tem-se que *a posteriori* $\beta|\sigma^2$ tem distribuição normal com média $\hat{\beta}_1$ e matriz de variâncias covariâncias $\sigma^2 V$ ou seja, em termos das estatísticas usuais a que coduz o método dos mínimos quadrados aplicado aos n_1 vectores (X_i,Y_i), $i \in obs$,

$$\hat{\sigma}_1^2 = \sum_{obs} \frac{(Y_i - X_i\hat{\beta}_1)^2}{n_1 - q}$$

$$\hat{\beta}_1 = V\left[\sum_{obs} X_i^T Y_i\right]^{-1}$$

onde,

$$V = \left[\sum_{obs} X_i^T X_i\right]^{-1}.$$

Então o método de imputação para este modelo é dado por:

(A) Extracção de uma variável aleatória g^* obtendo σ^{*2} dado por (4.16)

(B) Extracção de q variáveis $Z_i \sim N(0,1)$ que vão constituir o vector $Z = [Z_1 \ \ldots \ Z_q]$ e cálculo de $\beta^* = \hat{\beta}_1 + \sigma^*[V]^{1/2}Z$, onde $[V]^{1/2}$ é a raiz quadrada de V obtida pela factorização de Cholesky.

(C) Extracção de n_0 [2] valores de Y_{mis} tais que,

$$Y_i^* = X_i\beta^* + z_i\sigma^*,$$

onde os n_0 desvios z_i com distribuição $N(0,1)$ são extraídos independentemente.

Assim, um novo valor imputado para Y_{mis} é iniciado pela extracção de um novo valor do parâmetro σ_*^2. Assim se se pretendem m imputações os passos (A), (B) e (C) repetem-se m vezes independentemente.

[2] $n = n_0 + n_1$

BIBLIOGRAFIA

Barnett, V.(1991): *Sample Survey Principles and Methods*, Edward Arnold, Londres.

Box, G.E.P. e Tiao, G.C.(1973):*Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading.

Cochran, W.G.(1977): *Sampling Technics*, J. Wiley & Sons, Inc., 3rd Ed., Nova Iorque.

Crespo, J.L.(1976): *Muestreo de Poblaciones Finitas Aplicado al Diseño de Encuestas,* Instituto Nacional de Estadistica, Madrid.

Crespo, M.T.(1993): *Técnicas de Amostragem*, CESD, Lisboa.

Durão, A.C.C.(1993): *O Efeito das Não Respostas na Qualidade dos Inquéritos (Inquéritos às Empresas de Construção e Obras Públicas)*, INE, Lisboa.

Efron, B.(1994): "Missing Data, Imputation, and the Bootstrap", *JASA*, Vol.89, nº 426.

Geisser, S.(1993): *Predictive Inference: An Introduction*, Chapman & Hall, Londres.

Grosbras, Jean-Marie(1987): *Methods Statistiques des Sondages*, Economica, Paris.

Hansen, M.H. e Hurwitz, W.N.(1946): "The Problem of Nonresponse in Sample Surveys", *JASA*, nº 41.

Hansen, M.H., Hurwitz, W.N. e Madow,W.G.(1953): "*Sample Survey Methods and Theory*", Vol.2, Theory, J. Wiley & Sons Inc., Nova Iorque.

Kalton, G. e Kasprzyk, D.(1982): "Imputing for Missing Survey Response", *ASA, Proceedings of the Section on Survey Research Methods*.

Kish, L.(1965): *Survey Sampling*, J. Wiley & Sons Inc., Nova Iorque.

Lemeshow, S.(1985): "Nonresponse (In Sample Surveys)" in *Enciclopedia of Statistical Science*, Vol. 6.

Lessler, J.T. e Kalsbeeck, W.D.(1992): *Nonsampling Error in Surveys*, J. Wiley & Sons Inc., Nova Iorque.

Levy, P.S. e Lemeshow, S.(1991): *Sampling of Populations - methods and applications*, 2nd Ed., J. Wiley & Sons, Inc., Nova Iorque.

Little, R.J. A., Rubin, D.B.(1987): *Statistical Analysis with Missing Data,* J. Wiley & Sons, Inc., Nova Iorque.

Murteira, B.J.F.(1988): *Estatística: Inferência e Decisão*, Imprensa Nacional-Casa da Moeda, Lisboa.

Murteira, B.J.F.(1990): *Probabilidades e Estatística,* Vol.2, 2ª Ed., McGraw-Hill, Lisboa.

Murteira, B.J.F.(1993): *Análise Exploratória de Dados*, McGraw-Hill, Lisboa.

Platek, R., Singh M.P. e Tremblay, V.(1977): "Adjustment for Nonresponse in Surveys", *Survey Methodology*, Vol 3.

Rao, P.S.R.S.(1983): "Callbacks, Folow-Ups and Repeated Telephone Calls", in *Incomplete Data in Sample Surveys*, Vol. 2, Theory and Bibliographies, Academic Press, Nova Iorque.

Royall, R.M.(1992): "The Model Based (Predicton) Aproach to Finite Population Sampling Theory", in *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, Ed. Gosh, M. e Pathak, P.K., Institut of Mathematical Statistics.

Royall, R.M. e Herson, J.(1973): "Robust Estimation from Finite Populations", *JASA*, nº 68.

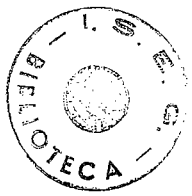
Rubin, D.B. e Little, R.J.A.(1987): *Statistical Analysis with Missing Data*, J. Wiley & Sons, Inc., Nova Iorque.

Rubin, D.B. (1987): *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Inc., Nova Iorque.

Rubin, D.B. e Schenker, N.(1986): "Multiple Imputation for Interval Estimation for Simple Random Samples With Ignorable Nonresponse", *JASA*, nº 81.

Särndal, Carl-Erik, Swesson, B. e Wretman J.,(1992): *Model Assisted Survey Sampling*, Springer-Verlag, Nova Iorque.

Selbold, J.(1988): "Survey Period Lenght, Unanswered Numbers, and Nonresponse in Telephone Surveys", in *Telephone survey Methodology*, J. Wiley & Sons, Inc., Nova Iorque.



Thompson, S.K. (1992): *Sampling*, J.Wiley & Sons, Inc., Nova Iorque.

Thomsen, Ib (1973): "A Note on the Efficency of Weighting Subclass Means to Reduce The Effects of Non-response When Analizing Survey Data", *Statistisk Tidskrift*, nº 4.

Tortora, R. (1985): "Nonsampling Errors in Surveys" in *Enciclopedia of Statistical Science*, Vol. 6.

Warner, S.L.(1965): "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias", *JASA*, nº 60.

Yates, F.(1949): *Sampling Methods for Censuses and Surveys*, C.Griffin & Co., Ltd., Londres.