VI CONGRESSO ANUAL

Tomar, 9 a 12 de Junho de 1998

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Tópicos de Sondagens

Paulo Gomes

TÓPICOS DE SONDAGEM

Paulo Gomes

ÍNDICE

Prefácio IV

Capítulo I - Generalidades

1 - Breves considerações sobre a história das sondagens 1
2 - Planeamento e execução de uma sondagem: principais etapas 3
3 - Métodos de sondagem 19

Capítulo II - Métodos empíricos de amostragem

1 - Introdução 21
2 - O métodos das quotas 21
3 - O métodos das unidades tipo 29

Capítulo III - Métodos probabilísticos

1 - Sondagem aleatória simples 30
2 - Sondagem aleatória estratificada 42
3 - Sondagem com probabilidades desiguais 55
4 - Sondagem em várias etapas 69

4.1 - Introdução 69
4.2 - Sondagem em conglomerados (cachos) 72
4.3 - Sondagem em duas etapas 77

Capítulo IV - Recurso à informação auxiliar com vista à melhoria de estimação

1 - O problema da recomposição da amostra 85
2 - Pós-estratificação 87
3 - Estimação do quociente 96
4 - Estimação do quociente na amostragem estratificada 104
5 - Estimação da regressão linear 109

Capítulo V - Estimação em pequenos domínios 114

Capítulo VI - Tratamento de não respostas

1 - Introdução 130
2 - Caracterização 132
3 - Extensão das não respostas 137
4 - Não respostas: avaliação empírica das consequências 143
5 - Utilização do mecanismo de respostas para a Estimação 149
6 - Modelização do mecanismo de respostas 156
7 - Avaliação através da Simulação de Monte Carlo 162
8 - Imputação de respostas - Estratégia na Face de Estimação 166

Capítulo VII - Tópicos sobre sondagens eleitorais

1 - Generalidades 172
2 - Reforma do Sistema Eleitoral Português 185

Anexos 200
Bibliografia 228

PREFÁCIO

Nos últimos cinquenta anos, a literatura relativa à inferência estatística desenvolveu-se em variadíssimos domínios e nas últimas décadas em áreas de importante aplicação prática tais como a Medicina, a Economia e o Ambiente. Paradoxalmente, apesar do incremento extraordinário das sondagens de opinião (onde as **sondagens eleitorais** e os **estudos de mercado** ocupam papel de relevo) e fundamentalmente das operações estatísticas por amostragem levadas a cabo pelos Institutos Oficiais de Estatística de todo o Mundo, a Comunidade Científica das Estatísticas Portuguesas tem revelado, com algumas boas excepções, um relativo desinteresse por este tema: tal como ocorreu na generalidade das Comunidades Científicas Internacionais, a ausência de um diálogo estreito entre Estatísticos, Estaticistas e mesmo especialistas da área das Ciências Sociais, não contribui para uma melhor compreensão dos importantes problemas que certas operações estatísticas suscitam, os quais, a serem identificados e analisados por essa comunidade, constituíram um bom exemplo do papel fundamental e imprescindível que a **"profissão de Estatístico"** pode e deve trazer ao desenvolvimento das Sociedades Contemporâneas.

Naturalmente que este tipo de comportamentos paradoxais ocorreram por **inércia dos Estatísticos**, por **inércia dos Estaticistas** e também por **inércia de alguns "peritos de sondagens"** com formação dominante na área das Ciências Sociais.

Contudo, constata-se em vários Países tais como o Canadá, Estados Unidos da América, Inglaterra e França, uma mudança radical de atitudes: os Teóricos da Estatística Matemática estão hoje plenamente convencidos que as sondagens proporcionam à reflexão e investigação problemas substanciais tanto ao nível da planificação de experiência e de modelização como no contexto propriamente dito de inferência.

O presente trabalho realizado a convite da Sociedade Portuguesa de Estatística e enquadrado num Curso sobre "Tópicos de Sondagem", que teremos a ocasião de apresentar, no VI Congresso da Sociedade Portuguesa de Estatística poderá constituir um pequeno passo para essa almejada articulação entre a classe de Investigadores Estatísticos Portugueses e as Instituições, Públicas e Privadas que desempenhem a tarefa quotidiana de implementar novos inquéritos.

Procurou-se alguma selectividade nos temas a incluir sem prejuízo de uma abordagem clássica das principais técnicas de amostragem a que recorrem as sondagens de opinião e as operações estatísticas realizadas no seio do Instituto Nacional de Estatística.

Não tivemos assim a preocupação de exaustividade na selecção dos temas aqui incluídos. Por outro lado, seleccionamos três temas que têm suscitado particular interesse nos últimos anos:

- **O tratamento de não respostas** - para o efeito convidamos o Dr. Osvaldo Caldeira (INE), o qual é responsável pela redacção do Capítulo VI onde esse tema é abordado.

- **A estimação em pequenos domínios**. O tema é tratado no Capítulo V por outro convidado, o Engº Pedro Coelho (ISEGI)

- e por último as **Sondagens Eleitorais,** as quais abordaremos genericamente, explicitando os principais problemas inerentes à sua realização para além de uma referência sucinta à **Reforma do Sistema Eleitoral Português**, convidando o Dr. Paulo Morais (Universidade Lusíada).

Manifestamos o nosso agradecimento ao Serviço de Metodologia do INE e em particular à responsável desse Serviço, Drª Helena Gomes, pela disponibilidade na preparação de alguns elementos detalhados da nova Metodologia do Inquérito ao Emprego (Série 98) que incluímos em anexo.

CAPÍTULO I - GENERALIDADES

1 - Breves considerações sobre a história das sondagens

As sondagens de opinião começam por surgir nos Estados Unidos ainda no século XIX a propósito de cobertura da Imprensa nas campanhas eleitorais Presidenciais, realizando-se então as primeiras sondagens pré-eleitorais junto dos eleitores, as quais foram objecto de grande apreço e curiosidade por parte da opinião pública.

Essas primeiras operações não incluíam qualquer critério de representatividade e alicerçavam-se em amostras de dimensão muito elevado (2 000 000 para o Literary Digest, em 1936).

Ora a data crucial na história das sondagens é sem dúvida o dia 3 de Novembro de 1936, dia da publicação dos resultados da eleição Presidencial nos Estados Unidos da América: com base no sondagem realizada pelo Literary Digest, onde se pergunta pelo voto entre dois candidatos (Lendon e Roosevelt), a revista conclui que Landon ganhara confortavelmente, indo de acordo ao sentimento generalizado dos Americanos.

Na mesma leitura, Georges Gallup interrogou apenas uma amostra de 4 000 Americanos **com perguntas sobre o voto em 1933 e o voto em 1936** e anunciou a eleição de Roosevelt em 56% dos votos. Ora Roosevelt foi reeleito com 60% dos votos!

De facto a amostra da revista Digest, apesar de corrigida de acordo com as variáveis sexo e idade, não traduzia a estrutura actual de voto.

Por outro lado, a amostra era claramente enviezada pois recorria a leitores da Revista!

Logo de seguida assiste-se ao aparecimento dos primeiros Institutos de Sondagem de Opinião Europeus, nomeadamente

em Inglaterra e em França; neste último País, Stoetzel, Professor da Sorbonne, cria o Instituto Francês de Opinião Pública e em 1947 ocorre, em Paris, o primeiro Congresso Internacional sobre toda esta temática.

Neyman (1891 - 1981) é provavelmente o nome mais sonante desta primeira fase de desenvolvimento das sondagens, evidenciando a superioridade das sondagens aleatórias em universos estratificados. Em 1938, Neyman introduz a amostragem em duas etapas e a utilização conjunta de funções de custo e de variância para o estabelecimento das partições óptimas do efectivo global de sondagem.

Surgem então duas vias essenciais de desenvolvimento: a primeira relacionada com problemas de selecção de amostras, estimação, tratamento de não respostas e a segunda, complemento da primeira, que enfatiza o problema do modelo probabilístico subjacente à teoria das sondagens.

Em 1943, Hansen e Hurwitz apresentam as primeiros resultados de estimação e de precisão dos estimadores quando a selecção dos indivíduos se realiza com probabilidades desiguais e no caso particular de planos de sondagem com reposição.

Hurwitz e Thompson generalizam em 1952 os resultados para o caso da extracção sem reposição e os principais algoritmos são desenvolvidos por Iates e Grundy em 1953.

As técnicas que recorrem a informação auxiliar são paralelamente desenvolvidas desde 1934 (com Neyman) com os estimadores do quociente, da diferença e de regressão e posteriormente objecto de investigação nos trabalhos de Cochran (1972), Hansen, Hurwitz e Madow (1953).

Por outro lado, no fim dos anos quarenta surgem os primeiros trabalhos sobre o problema das não respostas, totais ou parciais com os contributos de Hansen, Hurwitz (1946) e de Dalenius (1957).

O estudo de variabilidade dos estimadores constitui também um campo de investigação nomeadamente para estatísticas não lineares: desde os trabalhos de Malhalanobis (1946) até aos trabalhos sobre o Jackknife e mais recentemente o Bootstrap com Efron em 1979.
Nos últimos anos aprofundam-se as relações entre as sondagens e o modelo de estatística inferencial: mais geralmente, vários autores analisam a escolha dos estimadores e do plano de sondagem considerando que a população finita estudada é uma amostra extraída de uma superpopulação de dimensão infinita: destacam-se os trabalhos de Royall (1970), Sprolt (1971) e Särndall (1984).

2 - Planeamento e execução de uma sondagem: *Principais etapas*

A concepção e implementação de uma sondagem constitui, em geral, um processo pluridisciplinar de complexidade variável, exigindo a observância de um conjunto vasto de fases sequenciais das quais destacaremos:

2.1 - Formulação do problema em estudo e definição dos objectivos do inquérito

Esta primeira fase é porventura a mais exigente e mais importante de todo o processo pois compreende:

2.1.1 - O nível do conhecimento dos fenómenos que lhes estão adjacentes, respectivos conceitos e definições.

2.1.2 - A definição da população - alvo e da unidade amostral.

A definição da população não pode encerrar qualquer ambiguidade utilizando-se para o efeito variáveis que facilitem a delimitação desse universo tais como variáveis sócio-económicas, geográficas, assim como referências temporais. A unidade amostral pode ser elementar, isto é, reduzir-se a um só indivíduo ou ser constituída por um grupo de indivíduos (conglomerado), em função dos objectivos do estudo e da sua execução no terreno.
A título de exemplo, se pretendemos inquirir a população de uma freguesia, a unidade amostral pode ser o indivíduo, o conjunto dos indivíduos do mesmo agregado familiar ou todas as pessoas que habitam numa determinada zona da freguesia.

2.1.3 - A identificação da informação estatística que se pretende obter (e respectiva periodicidade se for caso disso), com referência à população - alvo, o que pressupõe uma primeira selecção das variáveis de inquirição (fase pré-questionário) a partir de um universo de variáveis directa ou indirectamente relacionadas com os objectivos do estudo. Relativamente ao tipo de informação recolhida e à dicotomia quantitativo-qualitativo persiste a "ditadura" do quantitativo sobre o qualitativo (quantofrenia) em parte na sequência da paradoxal credibilidade e objectividade atribuída ao quantitativo no "imaginário" dos utilizadores de informação.

2.1.4 - E, complementarmente a explicitação dos condicionalismos determinados pelos limites orçamentais e de precisão previamente fixados para a realização da operação estatística e pela disponibilidade de informação auxiliar potencialmente útil à compreensão do problema,

sugerindo hipóteses de trabalho, alimentando o rol de dados primários ou enquadrando o tratamento e a análise dos resultados do inquérito. Essa informação dita secundária pode ser encontrada em dados provenientes de dados administrativos, em referências bibliográficas, através da consulta a peritos e outros actores implicados ou pela análise de estudos similares já realizados noutras Instituições, noutros domínios afins ou noutros espaços geográficos. Contudo o recurso a este expediente exige uma prévia avaliação da qualidade das referidas fontes e da sua adaptabilidade ao novo contexto, implícito no problema agora em análise.

Note-se que este tipo de considerações reveste-se hoje da máxima importância se atendermos, por um lado, à necessidade de minimizarmos a dimensão da amostra, isto é o custo da sondagem para uma determinada precisão fixada por recurso a um processo amostral que beneficia à priori ou a posteriori de informação auxiliar e, por outro lado, à inevitabilidade em eliminarmos progressivamente toda e qualquer redundância que se traduza na recolha de informação já obtida parcialmente noutras operações estatísticas ou proveniente de actos administrativos facilmente adaptados para fins estatísticos: no caso dos inquéritos às empresas, trata-se de operacionalizarmos o recurso a informação já organizada para fins de gestão ou de declaração fiscal reduzindo assim a sobrecarga sobre as empresas, o custo da operação e o prazo para a disponibilização dos resultados do inquérito.

2.2 - Organização da base de sondagem (isto é, da lista das unidades de amostragem) que melhor se adapta ao problema em jogo no sentido de optimizar o binário custo - qualidade da sondagem

A construção ou acessibilidade a uma base de sondagem é muitas vezes complexa; no caso dos inquéritos junto das famílias, a listagem de alojamentos obtida pelo INE nos recenseamentos da população e habitação não podem, nos termos da lei actual ser disponibilizados a entidades externas, condicionando fortemente as opções de desenho amostral nas empresas privadas de sondagem.

Em todo o caso é dificil obtermos, em cada instante, uma base de sondagem completa ou que não contenha duplicações: a listagem de "telefones de residência" é um bom exemplo dessa situação exigindo cuidados suplementares na sua utilização. A base de sondagem que alimenta os inquéritos às empresas levados a cabo pelo INE está naturalmente sujeita ao processo de nascimento e morte das empresas e ao inevitável desfasamento parcial entre o momento em que se constitui a base de sondagem para um determinado inquérito de periodicidade anual e a conclusão do processo de actualização da base para esse ano de referência, com inclusão das novas empresas criadas e sujeitas a inquirição e exclusão das empresas extintas e outrora também sujeitas a essa inquirição.

Assim, em muitas situações, não só a população potencialmente inquirida não coincide exactamente com a população - alvo como a impossibilidade de recorrermos a uma base de sondagem completa e actualizada determina a constituição de uma sub-população chamada amostra-mãe obtida segundo um determinado desenho amostral isto é, constituindo-se uma primeira etapa desse desenho. Esta estratégia simplifica consideravelmente o processo pois apenas actualizamos uma parte "representativa" da população -alvo parte essa que assume o papel de base de sondagem para a segunda etapa do processo.

*Em Portugal o INE actualizava a amostra-mãe de 5 em 5 anos e a partir de 1998 recorrer-se-á a uma actualização parcial em cada ano.

2.3 - Concepção do questionário

Trata-se de uma etapa crucial de todo o processo. Em termos muito gerais enunciamos um conjunto de grandes princípios a observar na elaboração de cada questão (Droesbecke J., Lavallée P., 1998):

- Deve conter uma e uma só ideia
- Deve ser simples: recorrer a palavras simples e a uma linguagem acessível
- Deve ser clara e precisa (eliminando a possibilidade de interpretações subjectivas por parte do inquirido)
- Deve ser curta e directa (evitando as negações e sobretudo as duplas-negações)
- Deve ser lida e entendida facilmente
- Não deve sugerir uma resposta particular
- Não deve incluir elementos de emotividade
- Não deve contribuir para o surgimento de não-resposta: cuidados adicionais na formulação de "questões delicadas"

2.4 - Selecção da amostra

Estudaremos no capítulo seguinte uma variedade de "desenhos amostrais" ilustrando os contextos que justificam a sua escolha e comparando, quando possível a componente inferencial do erro da sondagem.
Sucintamente e em termos práticos recorre-se a uma amostra aleatória ou a uma amostra não aleatória ou ainda a um processo misto de amostragem sendo a selecção do desenho condicionada essencialmente pelos seguintes aspectos:

2.4.1 - A natureza dos dados disponíveis na população - alvo

Se podermos aceder a listas actualizadas dos indivíduos da população não estão excluídas abordagens do tipo selecção ao acaso ou a chamada "amostragem sistemática".

Exemplos:

*Listas eleitorais para a população com mais de 18 anos

*Listas dos lugares do país, disponíveis no Instituto Nacional de Estatística

*Lista das populações escolares num determinado grupo de escolas

2.4.2 - O grau de homogeneidade da população

Se a população não for "suficientemente homogénea", considerando o rol de questões presentes no questionário, é necessário recorrer à estratificação isto é, seleccionar um conjunto minimal de variáveis de estratificação ou seja

variáveis que determinem uma partição da população - alvo em grupos relativamente homogéneos chamados estratos. As variáveis de estratificação assimilam-se geralmente a um conjunto de critérios independentes cujo cruzamento, através das respectivas modalidades, determina o conjunto de estratos do estudo. Note-se que a homogeneidade referida é interpretada pelo sentido de atribuirmos exclusivamente às leis do acaso a variabilidade detectada no seio de uma amostra oriunda de um determinado estrato, relativamente a uma ou mais variáveis de interesse: Trata-se essencialmente de uma hipótese de trabalho, eventualmente já validada por estudos anteriores ou pelos objectivos intrínsecos ao nosso estudo (validação a priori) mas que está sempre sujeita a novas validações a posteriori através de procedimentos estatísticos confirmatórios adequados.

2.4.3 - A dispersão territorial do campo de observação

Se o território de observação for muito extenso ou se a população for muito dispersa, o modelo de selecção dos indivíduos ao acaso determinaria muitas deslocações dos entrevistadores, isto é, um custo elevado de sondagem. Em tais casos aconselha-se o recurso à chamada amostragem multi-etápica ou, em particular, a **amostragens em conglomerados** (cachos) ou do tipo areolar (para as grandes cidades). Num inquérito junto das famílias e abrangendo todo o território nacional, tratar-se-ia por exemplo de seleccionar, segundo um dado critério, uma amostra de lugares inquirindo numa segunda fase todas as famílias desse lugar.

2.4.4 - Os meios materiais disponíveis: Tempo, custo e logística

Os resultados de uma operação estatística valem em geral pela sua oportunidade o que nos obriga a estabelecer compromissos que determinam inexoravelmente desenhos amostrais sub-optimais mas capazes de responder aos aspectos essenciais do estudo em jogo.

A título de exemplo uma **amostragem em cotas** permite acelerar a recolha de informação pelo facto do inquiridor não estar sujeito a uma determinada lista de pessoas a interrogar, contornando até certo ponto o problema das recusas e das ausências. Evidentemente que esta opção exclui, em rigor, a possibilidade de explicitarmos uma medida do erro de amostragem além de outros inconvenientes, mas cuja enunciação não deverá significar, segundo a nossa opinião a censura a tais abordagens (correntemente utilizadas nas sondagens de opinião, realizadas pelas empresas privadas de sondagem), mas conduzir à inventariação de um conjunto de procedimentos cautelares que minimizem os efeitos negativos decorrentes dessa listagem de pontos fracos.

Complementarmente o factor custo é determinante na selecção do binário desenho amostral - método de recolha. Em sondagens de opinião que se repetem periodicamente é possível recorrermos, a montante do processo, a técnicas de classificação/análises factoriais indutoras de simplificações nesse desenho amostral. Adicionalmente o processo de selecção amostral pode ser influenciado pelo número de entrevistadores disponíveis e qualificados para a realização de um dado inquérito num dado intervalo de tempo.

Em conclusão, diríamos que uma mesma operação estatística pode contemplar variados "desenhos amostrais admissíveis", cada um deles acumulando, em geral, diferentes técnicas amostrais.

A análise comparativa da eficiência desses métodos na componente "erro de amostragem" não pode nem deve ser determinante já que outras considerações poderão justificar o recurso a abordagens alternativas susceptíveis de optimizar, condicionalmente um pacote de restrições iniciais, o chamado erro global de sondagem.

2.5 - Recolha de Informação

Qualquer projecto estatístico, quer seja recenseamento ou inquérito por amostragem é desenvolvido em várias fases. Uma dessas fases é a recolha da informação estatística, que pode ser feita por vários métodos:

- entrevista directa face a face com a utilização ou não de um computador portátil,
- entrevista por telefone;
- entrevista via postal;
- informação diária.

- **A entrevista directa face a face** é um processo de recolha de informação, em que o entrevistador administra um questionário a um entrevistado, na presença deste, e num período limitado de tempo. Este tipo de entrevista foi, durante muito tempo, o processo de recolha de informação mais usado e continua a ter muita importância na realização de alguns inquéritos. O aspecto personalizado do método pode estabelecer uma confiança entre o entrevistador e o entrevistado, que conduz ao sucesso da entrevista e à obtenção de dados de melhor qualidade.

- **A entrevista pelo telefone** é um processo de recolha de informação, em que o entrevistador administra um

questionário a um entrevistado, pelo telefone. É pouco usado no INE, porque muitos dos entrevistados não têm telefone.

- **A entrevista via postal** é um processo de recolha de informação, em que o questionário é enviado pelo correio para os respondentes, acompanhado por uma carta que explica os seus objectivos, e, pelas instruções de preenchimento, a que se junta um envelope já selado ou de porte pago, para que o entrevistado devolva o questionário preenchido.

- **A informação diária** difere dos outros tipos de entrevista já apresentados, porque o questionário é substituído por um diário onde os entrevistados deverão anotar, diariamente, a informação. O entrevistador tem que contactar o entrevistado, pelo menos duas vezes. Na primeira vez, o entrevistador explica o objectivo da informação diária, tenta conquistar a cooperação do entrevistado e explica o processo de recolha de dados. Na segunda vez, o entrevistador recolhe os dados, e, se estes não estão completos, ajuda o entrevistado a preencher o que está em falta. Durante o período de recolha podem existir mais visitas do entrevistador ao entrevistado, para lhe dar assistência, e, especialmente para o convencer a terminar a tarefa de preenchimento do diário. Este tipo de recolha é necessária para os inquéritos, que pressupõem acontecimentos muito frequentes e que apelam à memória dos entrevistados. Para evitar erros, que resultam do esquecimento dos factos, os entrevistados devem registar a informação, assim que os acontecimentos ocorrem.

Os recentes avanços tecnológicos deram origem ao aparecimento de métodos de recolha automática da informação, mais sofisticados do que os anteriores, que fazem

uso das novas tecnologias e que designam genericamente por entrevista assistida por computador.

Um dos métodos mais utilizado é o **CAPI** (**C**omputer **A**ssisted **P**ersonal **I**nterview). Começou a ser utilizado nos anos 80. É uma entrevista directa, face a face, assistida por computador. O entrevistador visita o entrevistado, lê-lhe as questões e anota as respostas directamente no computador portátil.

Este método tem vantagens relativamente ao método tradicional, em suporte de papel, porque:

- aumenta a flexibilidade na recolha da informação;
- diminui o tempo necessário à execução;
- reduz a intervenção manual que é passível de erros;
- padroniza os processos de entrevista

Implementação de uma sondagem: modelo organizativo e listagem das tarefas

Enunciaremos seguidamente uma lista detalhada das múltiplas tarefas subjacentes ao processo de desenvolvimento e implementação de uma sondagem associadas a cada uma das seguintes operações principais do processo:

1 - Objectivos

2 - Questionário

3 - Especificação do tratamento

4 - Sistema de tratamento

5 - Plano de sondagem (desenho amostral)

6 - Escolha da amostra

7 - Procedimentos de recolha de informação

8 - Trabalho de campo

9 - Tratamentos estatísticos dos dados

10 - Análise de resultados e relatório

Estas etapas principais são complementadas por outras actividades de "Planeamento e Controlo da Qualidade" tais como:

- Orçamento e calendário
- Documentação
- Formação
- Avaliação

pressupondo a existência de uma **equipa de projecto** estruturada segundo a seguinte lista de actividades:

A1 - Actividade de metodologia

A2 Actividade de "medida" (Questionário)

A3 - Trabalho de campo

A4 - Sistema informático

A5 - Actividade de tratamento

A6 - Actividade de coordenação

Listagem das tarefas (Fonte: Statistique Canada)

T1: Encontro com o cliente ou com a entidade(s) patrocinadora(s) da operação estatística ou com grupos previligiados de utilizadores

T2: Determinar os objectivos da sondagem

T3: Preparar o calendário global

T4: Preparar, para fins internos, o orçamento detalhado da sondagem

T5: Recolher informações disponíveis

T6: Preparar o glossário de conceitos

T7: Determinar as especificações dos quadros estatísticos

T8: Conceber o conteúdo do questionário

T9: Testar esse questionário mediante um inquérito-piloto

T10: Conceber a apresentação final do questionário

T11: Submeter o questionário a apreciações internas e externas

T12: Elaborar uma estratégia de codificação

T13: Preparar as especificações de verificação e de imputação

T14: Preparar as especificações de estimação

T15: Ensaiar o programa de estimação

T16: Preparar uma estratégia para a análise

T17: Elaborar um plano de amostragem

T18: Identificar as unidades de amostragem (base de sondagem)

T19: Redigir as especificações relativas à amostragem

T20: Escolher as unidades de amostragem

T21: Identificar as eventuais restrições informáticas

T22: Conceber a base de dados

T23: Elaborar e testar um programa de apuramento de dados

T24: Elaborar e testar um programa de validação

T25: Elaborar e testar um programa de imputação

T26: Elaborar e testar programas de cálculo de variância

T27: Administrar as tarefas informáticas

T28: Organizar a formação para a recepção e apuramento dos dados

T29: Conceber um programa de relações públicas com os utilizadores mais directamente implicados

T30: Definir os procedimentos de apoio à actividade de campo

T31: Determinar os procedimentos de controlo e preparar formulários de controlo para cada actividade de campo

T32: Preparar o manual de procedimentos para a actividade de campo

T33: Preparar a formação para cada actividade de campo

T34: Realizar a formação dos entrevistadores

T35: Realizar a recolha de dados

T36: Supervisionar cada uma das actividades de campo

T37: Proceder ao apuramento dos dados

T38: Produzir indicadores de qualidade dos dados

T39: Executar os programas de verificação, imputação e estimação

T40: Executar o programa de cálculo de variância dos estimadores

T41: Organizar os resultados da sondagem

T42: Realizar a análise dos resultados

T43: Preparar o relatório

T44: Avaliar globalmente a sondagem

T45: Submeter os resultados do inquérito à entidade cliente ou no caso das estatísticas oficiais proceder à divulgação pública dos resultados

Em conclusão, uma sondagem apresenta essencialmente as seguintes características:

- Uma complexidade variável
- Um carácter pluridisciplinar
- A necessidade de uma gestão rigorosa
- Grande selectividade nas escolhas metodológicas
- A preocupação da redução dos diferentes erros associados à operação estatística

3 - Métodos de Sondagem

1 - Introdução

Os diferentes métodos de sondagem agrupam-se em duas grandes categorias:

- os métodos **empíricos**
- os métodos **probabilísticos**

Por definição, as **sondagens probabilísticas** são tais que cada indivíduo da população tem uma determinada probabilidade, conhecida à priori, de pertencer à amostra (chamada **probabilidade de inclusão**).

Pelo contrário nas **sondagens empíricas** não é possível associar a cada indivíduo da população a probabilidade da sua inclusão na amostra.

Neste contexto, nas sondagens probabilísticas é possível, por recurso à teoria das probabilidades, estudar a **precisão** dos estimadores e por conseguinte controlar com rigor matemático a componente inferencial do erro global de uma sondagem. Esta situação explica a relativa desvalorização atribuída aos métodos empíricos, inspirados em considerações menos objectivas por vezes mesmo polémicas e geralmente classificados como menos rigorosos. Tais métodos não recorrem a **uma base de amostragem** fazendo contudo apelo a informação conhecida à priori ainda que não necessariamente actualizada (ex: recurso a informação censitária, particularmente desactualizada em determinados locais do País nos anos que antecedem a próxima operação censitária). A sua utilização assenta fundamentalmente em

considerações orçamentais e no factor tempo, tratando-se de uma abordagem frequentemente utilizada nas sondagens de opinião em inquéritos por entrevista directa ou por entrevista telefónica.

Observação: As estatísticas Oficiais Portuguesas não recorrem naturalmente aos métodos empíricos uma vez que os organismos produtores dispõem de uma base de sondagem para alicerçar as diversas operações estatísticas.

CAPÍTULO II - MÉTODOS EMPÍRICOS DE AMOSTRAGEM

1 - Introdução

Os métodos empíricos de sondagens caracterizam-se pelo facto de não ser possível determinar à priori a probabilidade de um determinado indivíduo pertencer à amostra e constituírem uma alternativa muito utilizada em sondagens de opinião e estudos de mercado onde, geralmente, não é **possível aceder a uma base de sondagem**.
Para remediar essa lacuna procura-se nas abordagens empíricas estabelecer um conjunto de orientações aos inquiridores de modo a reduzir os perigos de enviesamento decorrentes da arbitrariedade do entrevistador na selecção da amostra. Por outro lado, os métodos empíricos proporcionam processos de recolha da dados mais flexíveis tornando-se **mais rápidos** e **menos dispendiosos** que os processos de selecção probabilísticos, vantagens que explicam a sua grande divulgação.
Contudo apresentam como principal inconveniente o facto de **não permitirem**, em rigor, o **cálculo da precisão dos estimadores**.
Abordaremos aqui os dois tipos principais de sondagens empíricas**: o método das quotas** (incluindo a variante "itinerários") e os **métodos** das chamadas **"unidades-tipo".**

2 - O método das quotas

Segundo Vogt (1993) o método das quotas não é mais do que **uma amostra estratificada não aleatória**, isto é, construída a partir de uma partição da população em categorias e

posterior selecção de um certo número (quota) de elementos de cada uma das categorias segundo um critério não aleatório. A essência deste método reside no facto da **estrutura da amostra** reproduzir a **estrutura da população** segundo um conjunto de critérios estabelecidos e justificados à priori. Ou seja a filosofia deste método assenta nesta ideia-base: admitindo que os diferentes caracteres que se podem observar numa população não são independentes entre si, uma amostra que seja uma foto-redução da população no que diz respeito à distribuição estatística de alguns caracteres importantes, será também pouco diferente da população no que se relaciona com a distribuição dos caracteres não controlados.
Em suma o método das quotas resume-se a construir uma amostra que seja uma "**maquete da população**", isto é, um **modelo reduzido, segundo uma perspectiva principal, da população em estudo.**

Admitamos que se pretendia realizar um inquérito à população Portuguesa com mais de 18 anos.
Se, por exemplo, a população Portuguesa em 1998 é (segundo as estimativas do Gabinete de Estudos Demográficos do INE) tal que 48% da população com mais de 18 anos são homens e 52% são mulheres a amostra de quotas deveria incluir 48% de homens e 52% de mulheres.
Nas sondagens de opinião, onde o método das quotas é frequentemente utilizado, a escolha das "variáveis de estratificação" é feita geralmente tomando como base os dados do recenseamento da população. Por outro lado a selecção dessas variáveis segue a ideia dominante da estratificação no sentido de se maximizar a variabilidade entre os grupos definidos pelas quotas e consequentemente minimizar-se a variabilidade no interior desses grupos.

Podemos proceder a uma categorização unidimensional (por exemplo, sexo), bi-dimensional ou de um modo geral multidimensional construíndo assim uma amostra representativa da população segundo esses critérios tomados separadamente ou interrelacionados.

A folha geral de quotas, isto é, a especificação global do trabalho de campo pode pois resultar de quotas marginais ou de quotas cruzadas: no 1° caso facilita-se a tarefa do entrevistador já que para cada indivíduo a inquirir o entrevistador terá que obedecer a um só critério; no entanto este método tende a não assegurar a desejável representatividade da amostra. Não tem muito sentido falar em "amostra representativa da população" senão em termos de uma ou mais variáveis de quota previamente estabelecidas, situação que deve ser explicitada na divulgação dos resultados.

No 2° caso se for conhecida a estrutura cruzada da população, tudo se passa como se as quotas fossem determinadas por uma só variável (P. Ardilly, 1994); os entrevistadores terão que cumprir uma tarefa específica onde os indivíduos a entrevistar são distribuídos por diversos factores de estratificação tais como idade, sexo, grupo socio-económico, restringindo assim a liberdade do entrevistador quanto à escolha dos indivíduos a incorporar na amostra.

Naturalmente que esta abordagem pode tornar-se mais dispendiosa e morosa e dependente do bom desempenho do entrevistador em cumprir com rigor as especificações descritas na folha de quotas.

Exemplos:

1) Ilustra-se aqui a ficha técnica (resumida) de uma sondagem realizada pela Euroteste visando auscultar a tendência do eleitorado algumas

semanas antes das eleições legislativas de Julho de 1991:

Ficha Técnica

Universo: Indivíduos com idade igual ou superior a 18 anos residentes no Continente, em localidades com 10 ou mais fogos.

Dimensão da amostra: 902 entrevistas divididas em 424 homens e 478 mulheres.

Método de construção da amostra: Amostragem aleatória com estratificação por região e habitat para selecção do lar. Para a selecção dos indivíduos no lar foi utilizado o método por quotas sendo as variáveis de controlo: sexo, idade e grupo profissional.

Técnica de inquirição: Entrevista pessoal em cada do entrevistado mediante questionário estruturado.

Trabalho de campo: Realizado de 21 de Junho a 6 de Julho de 1991 por 36 entrevistadores.

Critério de Estratificação:Região/Habitat, Região/Ocupação, Sexo/Idade e Classe Social. No sentido de eliminar as pequenas diferenças que surgiram entre a amostra e o universo a que se reporta. No final a distribuição geográfica da amostra é a seguinte:

Grande Lisboa - 22,4%

Grande Porto - 10,3%

Litoral - 39,3%

Interior Norte - 15,2%

Interior Sul - 12,8%

2) Este exemplo refere-se a uma sondagem realizada em França e encomendada pelo Jornal Midi-Pyrénnés, um pouco antes das eleições legislativas de 1981. A cada entrevistador é atribuído um plano de trabalho que incluía uma folha de quotas do seguinte tipo:

Nome do Entrevistador: **Entrevistas a realizar:**		
Entrevistas	20	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Sexo H	10	1 2 3 4 5 6 7 8 9 10
M	10	1 2 3 4 5 6 7 8 9 10
Idade 18-34	9	1 2 3 4 5 6 7 8 9
35-49	5	1 2 3 4 5
50-64	4	1 2 3 4
+65	2	1 2
C.S.P.* do cabeça do casal Operário	1	1
Comerciante	1	1
Industrial-Profissão Liberal	1	1
Quadro Superior	1	1
Quadro Médio Empregado	7	1 2 3 4 5 6 7
Outro Trabalhador por conta de outrém	7	1 2 3 4 5 6 7
Inactivo-reformado	2	1 2

*** CSP - Categoria Socio-Profissional**

Importa pois que no final do inquérito a amostra recolhida obedeça às proporções previamente fixadas, baseadas na distribuição das características de controlo ao nível da população.

O plano de uma sondagem por quotas combina de um modo geral métodos aleatórios e não aleatórios, usualmente estabelece-se em 1° lugar um critério geográfico para a repartição da amostra cruzando regiões NUTS II (Norte, Centro e Lisboa, Vale do Tejo, Alentejo e Algarve, para o caso do Continente) com freguesias de dimensão populacional acima de um dado patamar mínimo e segundo um certo número de residentes em cada estrato.

Posteriormente em cada estrato considera-se uma amostragem em duas etapas: na 1ª, pode seleccionar-se uma amostra de freguesias proporcionalmente à dimensão populacional de cada freguesia, seleccionando numa 2ª etapa uma amostra **constante** e não aleatória segundo o método das quotas.

Estas considerações evidenciam que este plano de sondagem daria a cada indivíduo a mesma probabilidade de ser interrogado se fosse legítimo assimilar o método das quotas usado na última etapa a um processo de selecção equiprovável.

A este propósito os Institutos de sondagem entregam ao entrevistador não só a folha de quotas mas também um conjunto de instruções tais como:

- não entrevistar mais do que um indivíduo no mesmo imóvel;
- repartir os indivíduos a inquirir numa dada área geográfica por vários quarteirões ou lugares;
- inquirir em momentos diferentes do dia;
- não entrevistar indivíduos conhecidos.

Os quais visam reproduzir o melhor possível um processo equiprovável*.

Observações:

1) A "performance" do método das quotas depende, em 1º lugar, da possibilidade efectiva de incluirmos nos critérios de quotas as principais variáveis explicativas do fenómeno em estudo.
Em teoria é pois inevitável evitar o enviesamento implícito na selecção dos critérios pois a complexidade dos fenómenos não é em geral susceptível de ser explicada por um número reduzido de critérios. Note-se, no entanto, que no contexto das sondagens de opinião e em termos muito práticos os três critérios: idade, sexo, e categoria socio-profissional - têm-se revelado capazes de explicar a maioria dos comportamentos individuais.

2) Baseando-se no recurso a informação auxiliar os resultados provenientes deste método podem ser prejudicados pela inexistência de dados actualizados nas variáveis auxiliares tomadas como critérios o que pode aconselhar o recurso a quotas marginais, evitando dessa forma a imposição de quotas associadas a critérios cruzados cuja estrutura na população pode ter sofrido variações ao longo dos anos.

3) A importância atribuída no método das quotas ao trabalho realizado pelo entrevistador sugere uma especial atenção ao processo de controlo de qualidade à posteriori desse trabalho através de uma re-inquirição de uma dada

---

* Numa sondagem prática este tipo de considerações não legitima qualquer explicitação de erros de sondagem num processo de quotas.

percentagem de entrevistas via postal, telefone ou contacto directo.

Por vezes recorre-se também a métodos de estatística descritiva multidimensional tais como a análise factorial das correspondências múltiplas para realizar esse controlo já que, por essa via, é mais fácil detectar "casos atípicos", objecto de particular incidência em fase de re-inquirição.

4) No método de quotas, a tarefa essencial dos entrevistadores é cumprir as quotas estabelecidas, obedecendo ao conjunto de informações gerais estabelecidas. Poderá parecer à primeira vista que este processo não dá lugar a "não respostas", o que não traduz de facto a realidade já que durante a inquirição podem surgir indivíduos que se recusam a responder - este é mais um ponto fraco desta abordagem uma vez que não conhecemos o perfil desses não respondentes.

5) Estando a selecção final dos indivíduos sujeita à arbitrariedade dos entrevistadores os quais terão tendência a simplificar o seu trabalho e incluírem elementos de apreciação subjectiva na selecção, adopta-se frequentemente o método randim-route ou método dos itinerários aleatórios para evitar todo este tipo de inconvenientes e definir critérios objectivos de movimentação do entrevistador no terreno.
Este método não probabilístico é particularmente utilizado em trabalhos realizados em cidades ou localidades: o entrevistador é instruído a realizar o seu trabalho circunscrevendo-se a determinada área ou seguindo desde um determinado ponto de partida (seleccionado aleatoriamente a partir de um mapa ou de cartografia disponível) um certo itinerário na escolha das unidades a inquirir cobrindo assim o

melhor possível os quarteirões e ruas da área geográfica em questão (J. Antoine, 1990).
O principal inconveniente desta alternativa é o custo associado à preparação da operação.

3 - MÉTODO DAS UNIDADES TIPO

Trata-se de uma abordagem muito empírica que consiste em escolher um indivíduo médio considerado representativo de um grupo de indivíduos com características comuns. A ideia-base desta abordagem reside em admitirmos que os indivíduos similares ao **indivíduo-médio** no contexto das variáveis auxiliares que conduziram a esse indivíduo-médio serão também similares ao indivíduo -médio para as variáveis determinantes do inquérito ou seja admite-se entre as variáveis em análise e as variáveis auxiliares que alicerçam o indivíduo-médio, sendo essa relação constante para todos do indivíduos de um mesmo grupo (P. Ardilly, 1994).
Nas sondagens eleitorais recorre-se ainda, para fins de previsão de resultados finais, ao comportamento eleitoral em freguesias-tipo isto é, freguesias que se mantenham estáveis em vários períodos eleitorais no sentido de espelharem a selecção de forças entre os partidos à escala nacional.

CAPÍTULO III - MÉTODOS PROBABILÍSTICOS

III.1 - Sondagem aleatória simples (SAS): um modelo de referência.

1.1 - Introdução

Trata-se do modelo mais elementar de sondagem sem reposição onde numa determinada população com N indivíduos todas as amostras de dimensão n são equiprováveis com probabilidade igual a $1/{}^{N}C_{n}$. Por conseguinte é imediato verificar que a **probabilidade de um qualquer indivíduo pertencer à amostra é igual a *n/N*.**

Note-se que a sondagem aleatória simples constitui de facto um modelo de referência:

- com a qual comparamos os diferentes métodos de sondagem
- interveniente em certas fases de planos de sondagem mais complexos.

Por outro lado, a sondagem aleatória simples tem a particular propriedade **de não necessitar de informação auxiliar** relativa aos indivíduos da população - alvo; esta característica representa simultaneamente um ponto forte e um ponto fraco de análise:

- um **ponto forte** na medida em que propicia uma redução de custos
- um **ponto fraco** já que nos privamos de um expediente capaz de melhorar a precisão.

Observação:
Sendo π_i a probabilidade de inclusão do indivíduo i na amostra s, vimos que numa amostra aleatória simples:

$$\pi_i = \frac{n}{N} \qquad \text{para} \quad i=1, 2,..., N$$

logo

$$\sum_{i=1}^{N} \pi_i = N\left(\frac{n}{N}\right) = n$$

É fácil verificar que este resultado pode generalizar-se a qualquer outro tipo de sondagem de dimensão n:

"*Qualquer que seja o método de sondagem* $\sum_{i=1}^{N} \pi_i = n$ "

A verificação deste resultado é trivial se tomarmos uma variável indicatriz δ_i.

Tal que $\delta_i \begin{cases} 1 & \text{se } i \in s \\ 0 & \text{se } i \notin s \end{cases}$

δ_i é uma variável aleatória de Bernoulli tal que $P(\delta_i = 1) = \pi_i = E(\delta_i)$

Sendo $\sum_{i=1}^{N} \delta_i = n$

Resulta que $\sum_{i=1}^{N} E(\delta_i) = n$ ou $\sum_{i=1}^{N} \pi_i = n$

O conhecimento das probabilidades de inclusão π_i permitem-nos propor um estimador linear centrado para o parâmetro $T = \sum_{i=1}^{N} X_i$ onde X representa a variável em estudo:

Com efeito T é estimado por:

$$\hat{T} = \sum_{i \in s} \frac{X_i}{\pi_i} \qquad \text{ou} \qquad \sum_{i \in s} W_i X_i$$

onde W_i representa a ponderação atribuída ao indivíduo i, isto é, cada indivíduo da amostra **"representa"** um certo nº de indivíduos da população.

Se escrevermos $\hat{T}$ na forma:

$$\hat{T} = \sum_{i=1}^{N} \frac{X_i}{\pi_i} \delta_i$$

onde

$$\delta_i = \begin{cases} 1 & \text{se } i \in s \\ 0 & \text{se } i \notin s \end{cases}$$

$$E(\hat{T}) = \sum_{i=1}^{N} \frac{X_i}{\pi_i} E(\delta_i) = \sum_{i=1}^{N} X_i = T$$

o que prova, com este nível de generalidade, que $\hat{T}$ é um estimador centrado do parâmetro T.

No caso particular de uma amostragem aleatória simples $\pi_i = \frac{n}{N}$ resulta:

$$\hat{T} = \sum_{i \in s} \frac{N X_i}{n} = N\overline{X}$$

Neste contexto $W_i = \frac{1}{\pi_i} = \frac{N}{n}$: a ponderação é independente do indivíduo i seleccionado e da amostra s tomada; cada indivíduo da amostra representa $\frac{N}{n}$ indivíduos da população.

Representando $\mu = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{T}{N}$

resulta $\hat{\mu} = \frac{\hat{T}}{N} = \frac{1}{n}\sum_{i \in s} X_i = \overline{X}$

1.2 - Variância dos estimadores do total e da média numa amostragem aleatória simples.

$$\operatorname{var}(\hat{T}) = \operatorname{var}\left(\sum_{i=1}^{N} \frac{NX_i}{n}\delta_i\right)$$

$$= \frac{N^2}{n^2}\operatorname{var}\left(\sum_{i=1}^{N} X_i\delta_i\right)$$

$$= N^2\left(\frac{N-n}{N}\right)\frac{\sigma'^2}{n} = N^2(1-f)\frac{\sigma'^2}{n}$$

onde $\sigma'^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)^2$

(consultar anexo 1)

Consequentemente:

$$\text{var}(\hat{\mu}) = \text{var}\left(\frac{\hat{T}}{N}\right)^2$$

$$= (1-f)\frac{\sigma'^2}{n}$$

Observações: O resultado anterior revela que:

1) a precisão do estimador $\hat{\mu}$, avaliada pela *var* ($\hat{\mu}$), decresce à medida que *n* cresce;

2) essa precisão é dependente de σ'^2, isto é, do grau de "homogeneidade da população";

3) a precisão de $\hat{\mu}$ aumenta à medida que a taxa de sondagem $\left(\frac{n}{N}\right)$ tende para1;

4) podemos escrever *var* ($\hat{\mu}$) sob a forma:

$$\text{var}(\hat{\mu}) = \left(1-\frac{n}{N}\right)\frac{N\sigma^2}{N-1}\cdot\frac{1}{n}$$

$$= \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n} \qquad c/\ \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - u)^2$$

Se N for suficientemente grande comparado com n, $\frac{N-n}{N-1} \approx 1$ e $\text{var}(\hat{\mu}) \approx \frac{\sigma^2}{n}$, isto é, a precisão do estimador $\hat{\mu}$ numa sondagem aleatória simples é na generalidade dos casos práticos independente da dimensão *N* da população;

5) podemos avaliar o efeito de redução da $var\,(\hat{\mu})$ proveniente de uma extracção sem reposição quando comparada com uma extracção com reposição, com base em efectivos iguais:

$$\frac{var(\hat{\mu})_{\text{sem Rep}}}{var(\hat{\mu})_{\text{com Rep}}} = \frac{N-n}{N-1}$$

A este rácio chama-se efeito de sondagem (design effect)

1.3 - Estimadores das variâncias

Nas expressões de $\text{var}(\hat{T})$ e $var\,(\hat{\mu})$, σ'^2 não é conhecido. Tomaremos para estimador de σ'^2 a v.a..

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

pelo que

$$\widehat{var(\hat{T})} = N^2(1-f).\frac{S'^2}{n}$$

e

$$\widehat{var(\hat{\mu})} = (1-f).\frac{S'^2}{n}$$

1.4 - O caso particular das proporções

Podemos estar interessados em estimar a proporção de indivíduos num determinado domínio *D*:

para cada indivíduo $i \in D$ podemos então associar uma v.a. X_i de Bernoulli tal que:

$$X_i = \begin{cases} 1 & \text{se } i \in s \\ 0 & \text{se } i \notin s \end{cases}$$

então a proporção p de indivíduos da população que pertencem ao domínio D será estimado por:

$$\hat{p} = \sum_{i \in s} \frac{X_i}{n}$$

É obvio que $\hat{p}$ é um estimador centrado de p (trata-se afinal do estimador de uma média!) pelo que:

$$\text{var}(\hat{p}) = (1-f)\frac{\sigma'^2}{n}$$

Prova-se facilmente que $\sigma'^2 = \frac{N}{N-1} p(1-p)$; para N grande vem:

$$\text{var}(\hat{p}) = (1-f)\frac{p(1-p)}{n}$$

Para n>100 e N elevado toma-se:

$$\widehat{\text{var}(\hat{p})} \approx \frac{\hat{p}(1-\hat{p})}{n}$$

Se pretendermos estimar o n° N_D de indivíduos que pertencem ao domínio D basta tomar:

$$N_D = Np$$

donde, $$\hat{N}_D = N\hat{p}$$

Sendo para N grande $\text{var}\left(\hat{N}_D\right) = N^2\left(1-f\right)\dfrac{p\left(1-p\right)}{n}$

1.5 - O problema da dimensão da amostra numa sondagem aleatória simples

Para n suficientemente elevado, é geralmente possível construir um intervalo $(l_1,\ l_2)$ contendo o parâmetro θ desconhecido com uma dada probabilidade designada por $100(1 - \alpha)\%$ onde $1 - \alpha$ é geralmente elevado (0.95, 0.99, ...):

$$P\left(l_1 \le \theta \le l_2\right) = 100(1-\alpha)\%$$

No caso $\theta = \mu$ e para $1 - \alpha = 0.95$ obtém-se $\hat{\mu} \pm 1.96\sqrt{\text{var}\left(\hat{\mu}\right)}$

No caso $\theta = p$, proporção de indivíduos de população que pertence a um dado domínio D obtém-se para $1 - \alpha = 0.95$ um intervalo de confiança com limites:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} \qquad \text{(admitindo } n \text{ elevado)}$$

Note-se que nesta situação (frequentemente utilizada em sondagens de opinião) a amplitude do intervalo de confiança depende da estimativa $\hat{p}$ sendo $\hat{p}\left(1-\hat{p}\right)$ máximo para $\hat{p} = 0.5$.

No caso mais desfavorável podemos analisar a semi-amplitude ε (erro absoluto) para diferentes valores de n e tomando um nível de confiança de 95%:

$\hat{p} = 0.5$	
efectivo de sondagem	**erro absoluto**
100	10%
400	5%
1 000	3%
1 600	2.5%
10 000	1%

(admitimos aqui $1 - f \approx 1$*)*

Para melhor compreendermos a influência de p (e em termos práticos de $\hat{p}$) sobre o efectivo de sondagem associado a um erro absoluto pré-fixado podemos observar o quadro seguinte (P. Ardilly, 1994).

ε \ $\hat{p}$	**0.05**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**
0.5%	7 600	14 400	25 600	33 600	38 400	40 000
1%	1 900	3 600	6 400	8 400	9 600	10 000
2%	475	900	1 600	2 100	2 400	2 500
3%	211	400	711	933	1 066	1 111
4%	119	225	400	525	600	625
5%	76	144	256	336	384	400

Contudo pode fazer mais sentido analisar o problema avaliando o **erro relativo** $\frac{\varepsilon}{\hat{\theta}}$, isto é, analisar a dimensão da amostra que assegura uma dada precisão relativa.

O seguinte quadro descreve o erro relativo em % para o estimador $\hat{p}$ de uma proporção (tomando um nível de confiança 0.95 e uma taxa de sondagem tal que $1 - f \approx 1$):

		$\hat{p}$			
n	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**
100	60	40	31	24	20
200	42	28	21	18	14
300	34	23	17	14	12
500	26	19	15	12	10
1 000	18	13	9	8	6
2 000	14	9	7	6	4
5 000	8	6	4	4	3
10 000	6	4	3	3	2

Suponhamos agora que fixamos os limites de erro para um determinado nível de confiança, isto é, a condição:

$$P\left(\left|\overline{X}-\mu\right| \le \varepsilon\right) = 100(1-\alpha)\%$$

ou seja

$$P\left(\overline{X}-\varepsilon \le \mu \le \overline{X}+\varepsilon\right) = 100(1-\alpha)\%$$

e pretendemos determinar a dimensão mínima da amostra que assegura essa precisão com esse nível de confiança o quantil de ordem $\alpha/2$ de uma $N(0,1)$, (estamos a supor aqui $n>30$)

tem-se $C\sqrt{\text{var}\left(\overline{X}\right)} = \varepsilon$

ou $C\sqrt{\frac{\sigma^2}{n}.\frac{N-n}{N-1}} = \varepsilon$

ou $n = \frac{N\sigma^2}{(N-1)\frac{\varepsilon^2}{C^2}+\sigma^2}$

$$\text{ou } n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \quad (*)$$

onde n_0 representa a dimensão correspondente a um erro absoluto ε numa sondagem aleatória com reposição:

$$C\sqrt{\frac{\sigma^2}{n}} = \varepsilon \Rightarrow \frac{C^2\sigma^2}{n} = \varepsilon^2 \Rightarrow n_0 = \frac{C^2\sigma^2}{\varepsilon^2}$$

em termos práticos tomaremos:

$$\widehat{\text{var}(\overline{X})} = \frac{\hat{\sigma}^2}{n} \cdot \frac{N-n}{N} \qquad \text{logo}$$

$$\text{resultará } n = \frac{N\hat{\sigma}^2}{\frac{N\varepsilon^2}{C^2} + \hat{\sigma}^2} \qquad \text{(isto é a dimensão mínima da amostra vem } n^* = I(n) + 1\text{)}$$

No caso particular de estimação de uma proporção p obter-se-ia facilmente:

$$n = \frac{N\left(\frac{\varepsilon^2}{C^2} + \hat{p}(1-\hat{p})\right)}{\frac{\varepsilon^2}{C^2}N + \hat{p}(1-\hat{p})}$$

No pressuposto frequente de N ser elevado tomaremos uma expressão simplificada:

$$n = \frac{\hat{p}(1-\hat{p})C^2}{\varepsilon^2} + 1$$

*Esta expressão é interessante pois evidencia, como era de esperar, o facto de $n \approx n_0$ quando N é suficientemente elevado.

III.2 - AMOSTRAGEM ESTRATIFICADA

2.1 - Introdução

Na amostragem aleatória simples tivemos ocasião de enfatizar a característica de não recorrermos a qualquer tipo de informação auxiliar para a concretização do plano de sondagem (recolha de informação + formulação do estimador $\hat{\mu}$). No entanto, no caso do estimador $\hat{\mu}$, verificamos que a **precisão** desse estimador era directamente dependente da dispersão na população da variável em estudo X : essa dispersão avaliada aqui pelo parâmetro σ^2 é, em geral, muito sensível a valores extremos mesmo que eles sejam pouco frequentes, isto é, podemos incrementar a eficácia de todo o processo se realizarmos uma partição na população em grupos mais homogéneos chamados **estratos** de tal forma que a variabilidade inicial, decomposta em variabilidade **intra-estrato** e variabilidade **inter-estrato**, seja essencialmente explicada na componente inter-estrato.

Considerando uma população Ω dividida em k estratos $\Omega_i (i = 1, \cdots, k)$

$$\text{onde} \begin{cases} \underset{i,j}{\forall} \quad \Omega_i \cap \Omega_j = \varnothing \qquad (i \neq j) \\ \bigcup_{i=1}^{k} \Omega_i = \Omega \end{cases}$$

e tais que N_i representa a dimensão do estrato Ω_i

e $N = \sum_{i=1}^{k} N_i$ representa a dimensão da população Ω

e sendo:

- μ_i a média da variável X em estudo no estrato Ω_i
- σ_i^2 a variância de X no estrato Ω_i

podemos escrever

$$\sigma_i^2 = \frac{1}{N_i}\sum_{j=1}^{N_i}\left(X_{ij}-\mu_i\right)^2$$

onde X_{ij} representa o valor da variável X na observação j do estrato i.

Donde é imediato que $\mu = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \frac{\sum_{i=1}^{k} N_i\mu_i}{N}$

e $$\sigma^2 = \sum_{i=1}^{k}\frac{N_i{\sigma_i}^2}{N} + \sum_{i=1}^{k}\frac{N_i}{N}\left(\mu_i-\mu\right)^2$$

Ora como $\sum_{i=1}^{k}\frac{N_i}{N} = 1$ podemos concluir de facto que σ^2 é a soma de 2 médias ponderadas:

$\sum_{i=1}^{k}\frac{N_i{\sigma_i}^2}{N} \rightarrow$ é uma média ponderada da dispersão no interior dos estratos: representa uma **dispersão intra-estratos.**

$\sum_{i=1}^{k} \frac{N_i}{N}(\mu_i - \mu)^2 \rightarrow$ é uma média ponderada dos quadrados dos desvios entre as médias nos estratos e a média global: representa uma **dispersão inter-estratos**.

Este resultado vem confirmar a nossa intuição pois evidencia que a dispersão intra-estratos é **tanto menor quanto maior** for a dispersão inter-estratos pois a soma é constante e igual a σ^2.

Partindo desta ideia, a estratificação da população consiste em agrupar os indivíduos em k "grupos homogéneos" e a **amostragem estratificada** traduz-se na selecção de k amostras independentes seleccionadas, segundo um dado desenho amostral, em cada um dos estratos.
Teoricamente poderíamos assim admitir um desenho distinto para cada estrato.
Na prática não se utilizam, em geral, desenhos amostrais diferentes em cada estrato com excepção de situações em que pode fazer sentido inquirir todos os indivíduos de um determinado estrato (por exemplo, o estrato de empresas de um determinado sector de actividade com volume de negócios ou nº de assalariados acima de um certo patamar).
Assim o efectivo global de sondagem obtém-se pela justaposição dos efectivos associados a cada uma das sub-amostrais oriundas de cada estrato.
Ao nível da **base de sondagem**, a estratificação resume-se à etiquetagem de cada unidade estatística de modo a definir sem ambiguidade cada um dos estratos **por recurso a informação auxiliar.**

2.2 - Estimação e cálculo de precisão

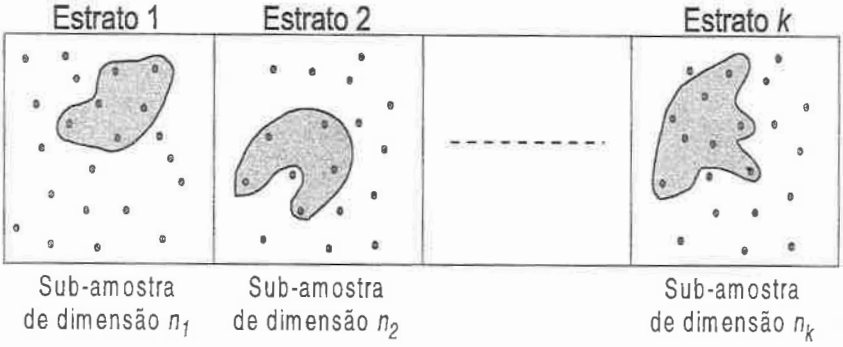
Admitindo uma amostragem aleatória simples em cada estrato, o estimador $\hat{\mu}$ vem naturalmente definido por:

$$\hat{\mu} = \sum_{i=1}^{k} \frac{N_i}{N} \hat{\mu}_i \qquad \text{onde} \qquad \hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{N_i} X_{ij} \delta_{ij}$$

$$\text{onde } \delta_{ij} = \begin{cases} 1 & \text{se} \quad j \in \text{ amostra do estrato } i \\ 0 & \text{se} \quad j \notin \text{ amostra do estrato } i \end{cases}$$

e sendo $E(\hat{\mu}_i) = \mu_i$ é imediato concluir que $E(\hat{\mu}) = \mu$

(supomos que os N_i são conhecidos para todo o i)



logo

$$\hat{\mu} = \sum_{i=1}^{k} \frac{N_i}{N} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \right)$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{N_i} \frac{N_i}{N n_i} X_{ij}$$

isto é, cada indivíduo j seleccionado no estrato i tem uma ponderação $W_i = \frac{N_i}{Nn_i}$, isto é, dependente do estrato, o que prova que a estratificação não é mais do que um método de selecção dos indivíduos com probabilidades desiguais (excepto casos particulares tais como a estratificação com dimensões amostrais proporcionais às dimensões dos estratos).

Por outro lado é imediato que

$$\begin{aligned} \operatorname{var}(\hat{\mu}) &= \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \operatorname{var}(\hat{\mu}_i) \\ &= \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \frac{\sigma_i^2}{n_i} \cdot \frac{N_i - 1}{N_i - 1} \end{aligned}$$

e

$$\widehat{\operatorname{var}(\hat{\mu})} = \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \frac{\hat{\sigma}_i^2}{n_i} \cdot \frac{N_i - 1}{N_i - 1}$$

onde

$$\hat{\sigma}_i^2 = \frac{1}{n-1} \sum_{j=1}^{n_i} \left(X_{ij} - \hat{\mu}_i\right)^2$$

Observação: se o valor de X_{ij} fosse conhecido para cada indivíduo, seria possível construir todos os estratos possíveis e seleccionar aqueles que minimizavam as variâncias em cada estrato: mas evidentemente que a inferência surge precisamente pelo facto de não acedermos aos valores da população na variável de interesse pelo que teremos que fazer apelo a uma ou mais variáveis auxiliares (particularmente

correlacionadas com X) construíndo os estratos em função dos valores assumidos na população pela ou pelas variáveis auxiliares.

2.3 - Comparação da eficiência de $\hat{\mu}$ numa amostragem estratificada e numa mostragem aleatória simples.

Seja $\hat{\mu}_{st}$ o estimador de μ numa amostragem estratificada:

$$\text{var}(\hat{\mu}_{st}) = \sum_{i=1}^{k} \frac{N_i^{\,2}}{N^2} \frac{\sigma_i'^{\,2}}{n_i} \frac{N_i - n_i}{N_i}$$

e $\hat{\mu}$ o estimador de μ numa amostragem aleatória simples:

$$\begin{aligned} \text{var}(\hat{\mu}) &= \frac{\sigma'^{\,2}}{n} \cdot \frac{N-n}{N} \\ &= \frac{N-n}{nN} \left(\sum_{i=1}^{k} \frac{N_i}{N} \sigma_i'^{\,2} + \sigma_b^2 \right) \end{aligned}$$

(onde σ_b^2 representa a variância inter-estratos)

então

$$\Delta = \text{var}(\hat{\mu}_{st}) - \text{var}(\hat{\mu})$$

$$= \underbrace{\sum_{i=1}^{k} \frac{N_i \sigma_i^2}{N^2} \left(\frac{N_i}{n_i} - \frac{N}{n} \right)}_{\mathbf{A}} - \underbrace{\frac{N-n}{nN} \sigma_b^2}_{\mathbf{B}}$$

Ora **B** é sempre ≥ 0 e zero sse $\sigma_b^2 = 0$ o que ocorre se $\mu_i = \mu$ para todo i . É obvio que **para** uma estratificação conveniente tal que σ_b^2 seja suficientemente elevado, **B** pode resultar superior a **A** isto é $\hat{\mu}_{st}$ mais eficiente que $\hat{\mu}$.
É importante o caso particular da amostragem proporcional onde $\frac{N_i}{N} = \frac{n_i}{n}$ para todo o i implicando **A**=0

$$\Rightarrow \Delta = -\frac{N-n}{nN}\sigma_b^2 \leq 0$$

isto é, a **amostragem estratificada proporcional conduz a um estimador $\hat{\mu}_{st}$ pelo menos tão eficiente que o estimador $\hat{\mu}$.**

Observação:

1) Poderíamos, através de um procedimento análogo, comparar a eficiência do estimador $\hat{\mu}_{st}$ numa amostragem com e sem reposição: como é de esperar o processo sem reposição é mais informativo logo mais eficiente.

2) A amostragem estratificada com dimensão em cada estrato proporcional à dimensão da sub-população Ω_i constitui um tipo de sondagem muito utilizado pela propriedade de eficiência relativa do estimador $\hat{\mu}$ já referido e pelo facto da abordagem não depender dos valores de X na população, isto é, qualquer conhecimento à priori sobre o comportamento dos indivíduos relativamente à variável X é irrelevante para a repartição do efectivo global de sondagem por cada estrato pois basta conhecer as dimensões $N_i (i = 1,2,\cdots,k)$.

2.4 - Repartição óptima do efectivo global de sondagem por cada estrato

Suponhamos que se dispõe de um determinado orçamento C^* para realizar um inquérito tendo-se, para o efeito, constituído um conjunto de k estratos avaliando o custo unitário C_i de sondagem em cada estrato e sendo C_0 os custos fixos de operação, a função **custo global** assume a forma:

$$C^* = C_0 + \sum_{i=1}^{k} n_i C_i$$

Pretende-se conhecer os efectivos de sondagem em cada estrato de forma a satisfazer a restrição orçamental.
A solução óptima poderá ser encontrada minimizando a $\text{var}(\hat{\mu})$ sujeita à condição

$$\sum_{i=1}^{k} n_i C_i = C_0^* \quad \text{com} \quad C_0^* = C^* - C_0$$

Trata-se de um problema de minimização condicionada de uma função k variáveis $(n_1, n_2, \cdots, n_k)$:

$$\text{var}(\hat{\mu}) = \sum_{i=1}^{k} \frac{N_i^2}{N^2} \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right)$$

A partir da função auxiliar de Lagrange

$$L(n_1, n_2, \cdots, n_k, \lambda) = \sum_{i=1}^{k} \frac{N_i^2}{N^2} \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right) + \lambda \left(\sum_{i=1}^{k} n_i C_i - C_0^* \right)$$

Obtém-se, supondo sem perda de generalidade, $\frac{N_i - n_i}{N_i - 1} \approx 1$ par todo o i

$$n_i = C_0 \frac{N_i \sigma_i / \sqrt{C_i}}{\sum_{i=1}^{k} N_i \sigma_i \sqrt{C_i}} = \gamma . \frac{N_i \sigma_i}{\sqrt{C_i}}$$

onde γ é uma constante de proporcionalidade, isto é, os n_i devem ser proporcionais a $\frac{N_i \sigma_i}{\sqrt{C_i}}$. O resultado obtido mostra que a amostra estratificada privilegia os estratos com N_i elevados, com dispersões σ_i^2 grandes e por fim os estratos cuja recolha é menos onerosa.

Observações:

1) Se eventualmente este resultado conduzir a valores de n_i superiores a N_i teríamos que impor a condição $n_i = N_i$ e retomar o processo integrando esta restrição desde o início do processo.

2) No caso particular em que o custo unitário é idêntico em todos os estratos, a restrição orçamental equivale a fixar o efectivo global n de sondagem e facilmente se obtém:

$$n_i = n \frac{N_i \sigma_i}{\sum_{i=1}^{k} N_i \sigma_i} \text{ (repartição óptima de Neyman)}$$

e a variância mínima do estimador $\hat{\mu}$ vem

$$\mathrm{var}(\hat{\mu})_{\min} = \frac{1}{n}\left(\sum_{i=1}^{k}\frac{N_i}{N}\sigma_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k}\frac{N_i}{N}\sigma_i^2$$

3) Facilmente se pode provar que sob uma dada restrição orçamental,

$$\mathrm{var}(\hat{\mu})_{\text{óptimo}} \leq \mathrm{var}(\hat{\mu})_{\text{estratificação proporcional}} \leq \mathrm{var}(\hat{\mu})_{\text{aleatória simples}}$$

4) Naturalmente que cada σ_i é desconhecido pelo que teremos que estimar **previamente** cada σ_i por recurso ao eventual conhecimento da dispersão de uma variável de estratificação relacionada linearmente com a variável de estudo ou aproveitando informação disponível em inquéritos similares precedentes. Finalmente **sugere-se** um procedimento ainda muito pouco utilizado nas empresas de sondagem e mesmo nos Institutos Nacionais de Estatística e que consiste em considerar um **desenho amostral em duas fases**, a primeira das quais tem por objectivo optimizar a segunda fase: na ocorrência poderíamos aproveitar a primeira fase para estimar as dispersões em cada um dos estratos.

5) Note-se que a repartição óptima de Neyman é competitiva, em termos de precisão final do estimador $\hat{\mu}$, relativamente à estratificação proporcional, no caso em que as dispersões σ_i são visivelmente diferentes ao longo dos estratos pois, caso contrário, as duas abordagens fornecem resultados muito similares.

6) Por outro lado e por oposição à estratificação proporcional, a repartição óptima de Neyman está directamente associada à estrutura da variável em estudo na população e segundo os k estratos ou seja a **optimalidade** subjacente a esta abordagem aplica-se apenas ao conjunto de variáveis particularmente correlacionadas com a **variável de referência** ou seja, esta abordagem é aconselhada em inquéritos monotemáticos onde tais considerações poderão ser relativamente plausíveis. Mas por vezes pretende-se realizar uma sondagem de opinião com múltiplos temas eventualmente destinados a diferentes clientes (os chamados inquéritos "*omnibus*") tornando-se em geral difícil que o plano de sondagem estabelecido (segundo orçamentos globais, em geral, demasiado restritivos) seja capaz de garantir níveis adequados de precisão em cada um desses temas.

7) Por vezes a estratificação parte, por justificações relacionadas com estudos anteriores, pela fixação de uma determinada proporção f_i de inquiridos em cada estrato sendo $\sum_{i=1}^{k} f_i = 1$, isto é, o efectivo de sondagem em cada estrato i é $n_i=nf_i$. E o problema consiste em determinar o efectivo global de sondagem n tal que $P\left(\left|\hat{\mu} - \mu\right| \leq \varepsilon\right) = 1 - \alpha$ (ou $P\left(\left|\hat{p} - p\right| \leq \varepsilon\right) = 1 - \alpha$).

Admitindo ε suficientemente pequeno tal que n_i é suficientemente elevado (n_i >50) e admitindo N_i - n_i > 50 facilmente se prova que:

$$n = \frac{\sum_{i=1}^{k} \left(N_i^2 A_i / f_i \right)}{N^2 D^2 + \sum_{i=1}^{k} N_i A_i}$$

onde $D = \left(\varepsilon / \chi_{\alpha/2} \right)$

e $A_i = p_i \left(1 - p_i \right)$ na estimação de p ou $A_i = \sigma_i^2$ na estimação de μ.

2.5 - Como constituir os estratos

Para a resolução deste problema há que identificar o conjunto de variáveis que melhor **discriminam** a população em estudo quer através do recurso à informação auxiliar relevante quer por via de métodos de análise estatística multivariada tais como métodos classificatórios (identificação de tipologias) de cariz exploratório e complementarmente abordagens factoriais explicatórias tais como análise factorial discriminante, geradoras de hipóteses de trabalho susceptíveis de posterior validação estatística através de abordagens confirmatórias.

8) Os efectivos de sondagem associados a cada estrato no sentido da partição óptima de Neyman minimizam a variância do estimador $\hat{\mu}$, isto é, do estimador correspondente à população global. Contudo pode ser também útil proceder a apuramentos em determinados estratos (caso, por exemplo, das estimações regionais) e porventura os n_i encontrados na situação global não satisfazem à "escala local", nomeadamente nos casos em que os N_i forem pequenos. Por

este facto é necessário encontrar soluções de compromisso conciliando as exigências "globais" com exigências "locais" .

A título de exemplo Bankier (Statistic Canada) desenvolveu um método baseado na optimização de um critério parametrizado o que conduz, numa situação limite, à repartição óptima de Neyman, mas viabilizando um compromisso que se traduz numa certa estabilidade do coeficiente de variação do estimador $\hat{\mu}_i$ em cada estrato i (P. Ardilly, 1994).

Seguidamente, coloca-se o problema do número de estratos a considerar no estudo: Se, por um lado, o aumento do número de estratos tende a favorecer a homogeneidade em cada estrato, por outro lado esse aumento determina um agravamento no custo do inquérito sendo evidente que um nº elevado de estrato para um efectivo global fixado determina pequenos efectivos de sondagem em cada estrato e consequentemente estimadores de σ_i^2 pouco precisos pondo em causa a qualidade da estimação de $\text{var}(\hat{\mu})$.

A este tipo de considerações adiciona-se o problema das não-respostas, particularmente com pequenos efectivos de sondagem.

Em conclusão, em termos práticos, sugere-se alguma contenção na definição dos estratos sem prejuízo do estabelecimento prévio de um nº mínimo de efectivo de sondagem em cada estrato.

III.3 - AMOSTRAGEM COM PROBABILIDADES DESIGUAIS

3.1 - Introdução

Na sondagem aleatória simples as probabilidades de inclusão na amostra são iguais para todos os indivíduos da população e iguais a *n/N*. Contudo esta abordagem não faz sentido em populações que englobam indivíduos (por exemplo, empresas) de dimensão muito diferente, já que se torna necessário tomar em consideração a importância relativa desses indivíduos no processo de selecção da amostra.
Neste contexto a recorrência a uma sondagem com probabilidades desiguais pode proporcionar uma melhoria na estimação do parâmetro associado à variável de interesse, na condição de **haver informação auxiliar** que justifique a probabilidade de inclusão não nula atribuída a cada indivíduo.

3.2 - Amostragem com probabilidades desiguais e com reposição

Consideraremos inicialmente a amostragem com **probabilidades desiguais e com reposição** utilizada por vezes no processo de selecção de "indivíduos" onde as dimensões das amostras são suficientemente elevadas para garantir uma boa aproximação aos resultados obtidos num processo **sem reposição**, para além da abordagem "com reposição" proporcionar grandes simplificações no cálculo da estimativa da variância do estimador em jogo.

Seja $K_i (i = 1,2,\cdots,n)$ a variável aleatória que representa o i-ésimo elemento seleccionado para a amostra. A *v.a.* K_i pode tomar os valores $1,2,\cdots,k,\cdots,N$

$$\text{sendo } P(K_i = k) = p_k \qquad \text{com } \sum_{k=1}^{N} p_k = 1$$

O estimador não enviesado de $T_X = \sum_{i=1}^{N} X_i$ (onde X representa a variável de interesse) é então definido por

$$\hat{T}_X = \frac{1}{n}\sum_{i=1}^{n} \frac{X_{K_i}}{p_{K_i}} \text{ (estimador de Hansen - Hurwitz)}$$

Nota: Concretizada a v.a. K_i : $K_i = k$
$(k = 1,\cdots,N)$

então $\begin{cases} X_{K_i} \equiv x_k \\ p_{K_i} \equiv p_k \end{cases}$
(notação simbólica)

isto é, o rácio $\frac{X_{K_i}}{p_{K_i}}$ toma, para cada i, os valores $\frac{x_k}{p_k}(k = 1,\cdots,N)$ com probabilidade p_k e $\hat{T}_X$ pode escrever-se na forma:

$$\hat{T}_X = \frac{1}{n}\sum_{i=1}^{n} Z_i \quad \text{com} \quad Z_i = \frac{X_{K_i}}{p_{K_i}} \Rightarrow E(\hat{T}_X) = \frac{1}{n}\sum_{i=1}^{n} E(Z_i)$$

$$E(Z_i) = \sum_{K=1}^{N} \frac{x_k}{p_k} P(K_i = k) = \sum_{K=1}^{N} X_k$$
$$= \mathrm{T}_X$$

$$\Rightarrow E(\hat{\mathrm{T}}_X) = \frac{1}{n} n \mathrm{T}_X = \mathrm{T}_X$$

A variância do estimador $\hat{\mathrm{T}}_X$ resulta imediatamente já que:

$$\mathrm{var}(\hat{\mathrm{T}}_X) = \frac{1}{n^2} \sum_{i=1}^{n} V(Z_i) = E\left[\frac{X_{K_i}}{p_{K_i}} - \mathrm{T}_X \right]^2$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{N} \left(\frac{x_k}{p_k} - \mathrm{T}_X \right) p_k}_{V_I} = \frac{V_I}{n}$$

e uma estimativa não enviesada para $V(\hat{\mathrm{T}}_X)$ é dada por:

$$\widehat{v(\hat{\mathrm{T}}_X)} = \frac{\hat{V}_I}{n}$$

$$\text{com } \hat{V}_I = \frac{1}{n-1} \sum_{K=1}^{n} \left(\frac{X_K}{p_K} - \hat{\mathrm{T}}_X \right)^2$$

$$\Rightarrow \widehat{V(\hat{\mathrm{T}}_X)} = \frac{1}{n(n-1)} \sum_{K=1}^{n} \left(\frac{X_K}{p_K} - \hat{\mathrm{T}}_x \right)^2$$

Observações:

1) Se as probabilidades p_K forem proporcionais aos valores da variável X_K então o rácio X_K / p_K resultaria constante e o estimador de Hansen-Hurwitz teria variância nula. No entanto, esta situação raramente pode ser utilizada na prática pois desconhecemos os valores da variável X. Mas o **conhecimento de uma outra variável auxiliar Y** aproximadamente proporcional a X conduz-nos a:

$$X_k = Cp_k$$

e $X_k \approx rY_k$

ou $$p_k \approx \frac{rY_k}{C} \Rightarrow \sum_{k=1}^{N} p_k \approx \frac{r}{c} \sum_{k=1}^{N} Y_k$$

$$\Rightarrow \frac{r}{c} \approx \frac{1}{\sum_{k=1}^{N} Y_k}$$

logo $$p_k \approx \frac{Y_k}{\sum_{k=1}^{N} Y_k}$$

Trata-se de um **desenho amostral probabilístico com reposição e com probabilidades proporcionais à dimensão.**

2) Podemos reforçar a análise da questão abordada na observação anterior, analisando em que condições.

$$V\left(\hat{\mathrm{T}}_X\right)_{prob} < V\left(\hat{\mathrm{T}}_X\right)_{aas}$$

$$V\left(\hat{\mathrm{T}}_X\right)_{prob} = \frac{1}{n}\sum_{k=1}^{N}\left(\frac{x_k}{p_k} - \mathrm{T}_x\right)^2 p_k$$

$$= \frac{1}{n}\left(\sum_{k=1}^{N}\frac{x_k^2}{p_k} - \left(\mathrm{T}_x\right)^2\right)$$

e $$V\left(\hat{\mathrm{T}}_X\right)_{aas} = N^2.\frac{\sigma^2}{n} = \frac{1}{n}\left(\sum_{k=1}^{N} Nx_k^2 - \left(\mathrm{T}_X\right)^2\right)$$

pelo que

$$V\left(\hat{\mathrm{T}}_X\right)_{prob} < V\left(\hat{\mathrm{T}}_X\right)_{aas}$$

sse

$$\sum_{k=1}^{N}\frac{X_k^2}{p_k} < \sum_{k=1}^{N} Nx_k^2$$

ou $$\sum_{k=1}^{N}\left(\frac{1}{p_k} - \frac{1}{1/N}\right)x_k^2 < 0$$

Ora esta expressão será tanto "mais negativa" quando para x_k^2 **grande** é satisfeita a condição $p_k > \frac{1}{N}$ e para x_k^2 **pequeno** for satisfeita a condição $p_k < \frac{1}{N}$, isto é, X_k correlacionado positivamente com p_k, o que está de acordo com a nossa intuição.

3.3 - Estimador de Horvitz - Thompson

3.3.1 - Introdução

Seja p_i a probabilidade de inclusão de um indivíduo i da população na amostra sem reposição e de dimensão n.

Já foi referido que

$$\sum_{i=1}^{N} p_i = \begin{cases} n & \text{se admitirmos uma extracção com dimensão amostral fixa.} \\ E(N) & \text{Caso a dimensão seja uma v.a. N.} \end{cases}$$

Admitamos que na sequência de um dado algoritmo, a selecção dos indivíduos é tal que podemos associar a cada um deles uma **probabilidade de inclusão conhecida** p_i.

O **estimador de Horvitz-Thompson ou estimador das somas dilatadas** é definido por:

$$\hat{T} = \sum_{i \in s} \frac{X_i}{p_i} \qquad (p_i > 0)$$

é imediato verificar que $E(\hat{T}) = T$ já que $\hat{T}$ pode assumir a forma:

$$\hat{T} = \sum_{i=1}^{N} \frac{x_i}{p_i} \varepsilon_i \qquad (p_i > 0)$$

onde ε_i é uma *v.a.* tal que:

$$\varepsilon_i = \begin{cases} 1 & \text{se} \quad i \in amostra \\ 0 & \text{se} \quad i \notin amostra \end{cases}$$

sendo $P(\varepsilon_i = 1) = P$ (inclusão do indivíduo i) $= p_i \Rightarrow E(\varepsilon_i) = p_i$

Observação:

1) Note-se que estamos a admitir que todas os indivíduos da população têm uma **probabilidade não nula** de pertencer à amostra: se p_i fosse nulo para certos indivíduos estaríamos a excluí-los à partida introduzindo o problema da "**não cobertura na base de sondagem**" pelo que $\hat{T}_X$ seria um estimador enviesado de T_X na população (apenas estimaria sem enviesamento o total de X para os indivíduos da população onde p_i seria positivo).

2) Na amostragem aleatória simples as probabilidades de inclusão p_i são todas iguais: $p_i = \frac{n}{N}$ para todo o i $\Rightarrow \hat{T}_X = \sum_{i \in s} \frac{X_i}{n/N} = N\overline{X}$ como era de esperar.

3.3.2 - Estudo da precisão do estimador

Admitindo a extracção da amostra numa só etapa, a variância do estimador $\hat{T}_X$ vem dado por:

$$V\left(\hat{T}_X\right) = \sum_{i=1}^{N} X_i^2 . \frac{1-p_i}{p_i} + \sum_{i=1}^{N}\sum_{j\neq 1}^{N} \frac{X_i}{p_i}\frac{X_j}{p_j}\left(p_{ij} - p_i p_j\right)$$

onde p_{ij} representa a probabilidade de inclusão simultânea dos indivíduos i e j na amostra s.
(ver anexo **3**)

É interessante considerar a $V\left(\hat{T}_X\right)$ sob a forma:

$$v\left(\hat{T}_X\right) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq 1}^{N}\left(p_i p_j - p_{ij}\right)\left(\frac{X_i}{p_i} - \frac{X_j}{p_j}\right)^2$$

Esta expressão mostra que a $v\left(\hat{T}_X\right)$ reduzir-se-ia a zero se

$$\left(\frac{X_i}{p_i} - \frac{X_j}{p_j}\right)^2 = 0 \quad \underset{(i,j)}{\forall}$$

isto é se as probabilidades de inclusão forem proporcionais a X_i. Esta constatação vale sobretudo pela indicação que fornece à priori, quanto ao nível de precisão do estimador $\hat{T}_X$ se as probabilidades de inclusão p_i forem aproximadamente proporcionais a X_i para cada indivíduo i da população.

No caso extremo onde $X_i = \lambda P_i$

$$T_X = \sum_{i=1}^{N} X_i = \sum_{i=1}^{N} \lambda P_i = n\lambda$$

e $$\hat{T}_X = \sum_{i\in s} \frac{X_i}{p_i} = \sum_{i\in s} \frac{\lambda P_i}{P_i} = n\lambda = T$$

como era de esperar.

Na prática recorre-se naturalmente a uma sondagem representativa segundo uma variável auxiliar Y_i conhecida ao nível da população.

No caso da uma **amostra de dimensão fixa *n*,** basta tomar

$$p_i = n.\frac{Y_i}{\sum_{i=1}^{N} Y_i}$$

Observações:

1) Este resultado exige que para todo o indivíduo *i* da população seja satisfeita a condição:

$$nY_i \leq \sum_{i=1}^{N} Y_i$$

por conseguinte se para um determinado indivíduo *i*, nY_i é superior $\sum_{i=1}^{N} Y_i$, podemos atribuir uma probabilidade um à sua inclusão na amostra retomando o processo proporcional para os *n-1* restantes e calculando $p_i = (n-1)\frac{Y_i}{\sum_{i=1}^{N-1} Y_i}$.

2)

a) Com toda a generalidade a $V(\hat{T}_X)$ é estimada pelo estimador:

$$\widehat{V_1(\hat{T}_X)} = \sum_{i \in s} \frac{1-p_i}{p_i^2} x_i^2 + \sum_{i \in s} \sum_{\substack{j \in s \\ j \neq i}} \frac{P_{ij} - P_i P_j}{P_i P_j P_{ij}} x_i x_j$$

onde, para todo o (i,j) é suposto $p_{ij} \neq 0$

b) No caso particular em que a amostra é considerada da dimensão fixa n podemos considerar um estimador concorrente (também centrado),

$$\widehat{V_2(\hat{T}_X)} = \frac{1}{2} \sum_{i \in s} \sum_{\substack{j \in s \\ j \neq i}} \frac{p_i p_j - p_{ij}}{p_{ij}} \left(\frac{x_i}{p_i} - \frac{x_j}{p_j} \right)^2$$

(P.Ardilly, 1994)

a) estes estimadores são diferentes mas ambos impõem que p_{ij} seja $\neq 0$ para todo (i,j) provando-se que caso contrário $\widehat{V_2(\hat{T}_X)}$ sub-estima $V(\hat{T}_X)$ e $\widehat{V_1(\hat{T}_X)}$ sobre-estima $V(\hat{T}_X)$. Por outro lado para além de dificuldade em conhecer os p_{ij} para cada (i,j) o cálculo torna-se complexo já que o número de parcelas que define a $\widehat{V(\hat{T}_X)}$ atinge rapidamente valores da ordem das centenas de milhar.

3) Este tipo de considerações sugere-nos algum pragmatismo quando equacionarmos o recurso a uma amostragem probabilística com probabilidades desiguais (e sem reposição):

Se temos disponível uma variável auxiliar Y aproximadamente proporcional a X e se os X_i apresentarem uma dispersão (potencial) elevada, então o recurso a uma amostragem probabilidade com p_i proporcionais a Y_i faz todo o sentido e tenderá a produzir resultados com boa precisão. No entanto se a variável em jogo apresentar diminuta variabilidade, provavelmente a sondagem aleatória simples poderá ser a escolha mais apropriada.

4) Não desenvolveremos neste curso abordagens de amostragem com probabilidades desiguais num processo multi-etápico. Tem especial interesse a situação em que se impõe, em cada etapa, dimensões amostrais constantes.

Por exemplo num inquérito junto das famílias é frequente pensar numa selecção de famílias em 3 etapas:

- na 1ª selecciona-se α freguesias proporcionalmente à sua dimensão populacional;
- na 2ª etapa tomam-se β secções estatísticas em cada uma das α freguesias, proporcionalmente à dimensão das secções;
- e na última etapa seleccionaram-se através de um esquema aleatórias simples (ou eventualmente através de um esquema sistemático) as γ famílias a inquirir em cada secção seleccionada (sobre este tema, pode consultar P.Ardilly, 1994).

Apresentaremos no **Anexo 4** algumas notas sobre a metodologia da nova série do inquérito ao emprego (1998).

5) A consideração de desenhos amostrais mais complexos envolvendo processos multi-etápicos e eventualmente estratificações numa ou mais dessas etapas, gera grandes complicações no cálculo da estimativa da

variância do estimador $\hat{T}_X$. Neste contexto a alternativa consiste em tomar uma amostragem com probabilidades desiguais em cada etapa mas **com reposição**, facilitando substancialmente o cálculo

da $\widehat{V(\hat{T}_X)}$, a qual constitui uma majorante da $\widehat{V(\hat{T}_X)}$ num processo sem reposição.

3.4 - Algoritmos para a selecção de amostra no contexto da amostragem com probabilidades desiguais.

Existem dezenas de algoritmos que asseguram a cada indivíduo da base de sondagem uma probabilidade de selecção p_i onde os p_i são conhecidos $0 < p_i \leq 1$.

Um dos principais obstáculos inerentes à extracção com probabilidades desiguais reside na dificuldade em calcular a probabilidade p_{ij} indispensável para os cálculos de variância do estimador, probabilidade essa que é **dependente** do algoritmo considerado.

Para solucionar este complexo problema pode recorrer-se a hipóteses simplificadoras ou utilizando apenas medidas aproximativas da variância do estimador.

Por outro lado pode obter-se soluções **impróprias** para a variância do estimador no caso de uma amostra de dimensão fixa pelo facto de serem negativas: basta que ocorra um número elevado de quantidades $\left(P_i P_j - P_{ij}\right)$ de sinal negativo.

Por esse facto os algoritmos admissíveis devem satisfazer as chamadas condições de Yates-Grundi:

$$\textbf{para todo o } i \textbf{ e todo } j(\neq i) \quad p_i p_j - p_{ij} \geq 0$$

Refira-se três algoritmos que reúnem simplicidade e eficácia prática e que são de uso particularmente frequente:

- **extracção de Bernoulli**
- **método sistemático**
- **método RHC (Rao, Hartley, Cochram)**

A **extracção de Bernoulli** é porventura o método cuja implementação é mais elementar (embora menos eficaz que os outros dois). O princípio do método consiste simplesmente a atribuir a cada indivíduo i de uma população com N indivíduos uma probabilidade de inclusão p_i e gerar um número pseudo-aleatório u a partir de uma lei uniforme entre 0 e 1.

Se $u_i < P_i$, toma-se o indivíduo i ou seja,

P (inclusão do indivíduo i na amostra)

$$= P(u_i < p_i) = p_i$$

Caso contrário i não é seleccionado. Nestas condições geramos um número u_i para cada indivíduo da população e independentemente de qualquer outro. Esta abordagem permite consequentemente calcular imediatamente p_{ij}:

$p_{ij} = p_i p_j$ atendendo à independência.

O grande inconveniente desta abordagem reside no facto da **dimensão da amostra ser uma v.a.** a qual pode mesmo tomar o valor zero!

Essa v.a. $N = \sum_{i=1}^{N} \delta_i$

onde δ_i é uma v.a. tal que:

$$\delta_i = \begin{cases} 1 & \text{se} \quad \mu_i < p_i \\ 0 & \text{senão} \end{cases}$$

$$P(\delta_i = 1) = P(\mu_i < p_i) = p_i \Rightarrow E(\delta_i) = p_i$$

$$\Rightarrow E(N) = \sum_{i=1}^{N} E(\delta_i) = \sum_{i=1}^{N} p_i$$

Neste caso pode provar-se que o estimador $\hat{T}_X = \sum_{i \in s} \frac{X_i}{p_i}$ admite uma variância igual a

$$v(\hat{T}_X) = \sum_{i=1}^{N} \frac{1 - P_i}{P_i} X_i^2$$

Não abordaremos aqui os outros dois métodos (consultar, p. ex., P. Ardilly)

III.4 - SONDAGENS EM VÁRIAS ETAPAS

4.1. - Introdução

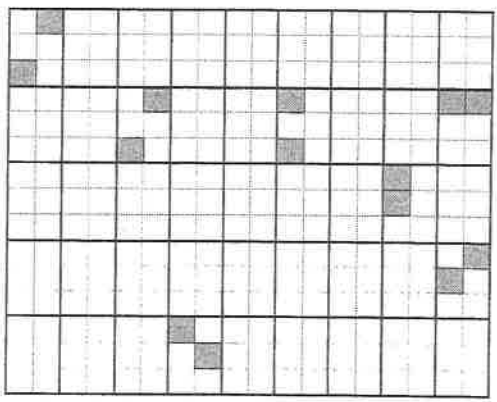
A amostragem aleatória simples apresenta em termos práticos duas grandes limitações:

1 - A necessidade de dispormos de uma **base de sondagem** completa, isto é, de uma base que identifique todos os indivíduos da população de referência

2 - O custo elevado de inquirição em inquéritos cujo âmbito geográfico tem alguma dimensão decorrente de um elevado custo de deslocação para efectivar uma dada entrevista directa.

Neste contexto podemos recorrer a abordagens alternativas do tipo sondagens multi-etápicas segundo as quais definimos numa 1ª etapa uma dada partição no conjunto dos indivíduos da população a inquirir seleccionando segundo um esquema aleatório simples (ou eventualmente por outros métodos) uma amostra de um certo número de grupos chamados **unidades primárias** ou seja cada grupo previamente constituído funciona como "unidade de amostragem" (se a escolha dessa amostra de unidades primárias obedecer a um esquema aleatório simples estamos a atribuir probabilidades iguais de inclusão o que só faz algum sentido se admitirmos que estamos perante uma população homogénea de grupos).
Numa 2ª etapa procedemos então à selecção de indivíduos de cada uma das m unidades primárias seleccionadas na 1ª etapa (novamente segundo uma amostragem aleatória simples ou outra) obtendo agora as chamadas **unidades secundárias**:

Como vemos só necessitamos do conhecimento da base de



(selecção de m=7 unidades primárias e 2 unidades secundárias em cada unidade primária)

sondagem de indivíduos associados às unidades primárias seleccionadas na 1ª etapa. Naturalmente que este processo pode generalizar-se a várias etapas.

Exemplo:

1) Sondagens eleitorais: As sondagens à boca das urnas constituem um bom exemplo de uma sondagem em 2 etapas: na 1ª seleccionam-se os "locais de voto" para onde se enviam entrevistadores e seguidamente toma-se uma amostra de eleitores em cada um desses locais.

2) Pretende-se realizar uma sondagem tendo em vista avaliar o nível de imunização contra o sarampo de crianças nas escolas primárias de Trás-os-Montes; podemos conceber um processo em 3 etapas:

   1ª etapa: Considerar uma amostra de concelhos (eventualmente segundo uma selecção

proporcional ao nº de crianças inscritas nas escolas primárias desse concelho.

2ª etapa: Retirar segundo um esquema idêntico uma determinada amostra de escolas em cada um dos concelhos previamente escolhidos.

3ª etapa: Considerar uma amostra de salas de aula em cada escola e finalmente analisar, por exemplo, todas as crianças em cada uma dessas salas.

3) Pretendia-se realizar uma sondagem numa cidade junto de agregados familiares sendo a selecção aleatória simples um método admissível tendo em conta a natureza específica do inquérito; para uma determinada precisão fixada sugere-se a inquirição de 600 casas dessa cidade.

Suponhamos então que considerações relacionadas com o custo/tempo de deslocação ou com a base de sondagem justificam o recurso a um processo alternativo consistindo na selecção prévia de um determinado nº de sub-secções estatísticas (quarteirões) de cidade inquirindo seguidamente uma amostra de lares em cada um dos quarteirões seleccionados. Tal procedimento alternativo **pode ser ou não** uma boa abordagem! Com efeito o perigo reside no facto de haver uma tendência de homogeneização de perfil socio-económico no interior do quarteirão a qual, a ser verdade, pode introduzir alguma redundância no processo de selecção de lares em cada unidade primária, e por conseguinte a natureza da ou das variáveis presentes no questionário deverá conduzir-nos a uma avaliação prévia do problema.

4.2 - UM CASO PARTICULAR DE AMOSTRAGEM BIETÁPICA: Amostragem em conglomerados (cachos).

Na amostragem em conglomerados processa-se a selecção de m unidades primárias (cachos) entre as M disponíveis e inquire-se todos os indivíduos de cada um desses cachos.

Exemplos:

1) Estudos sobre a prescrição médica.
   Considere-se uma amostra de médicos considerada como "cachos" de pacientes ou de prescrições.

2) Inquérito aos passageiros de uma determinada linha de uma Companhia Aérea: os cachos são os vôos seleccionados inquirindo-se no avião todos os passageiros.

Notações:

Seja N_i - dimensão do cacho i (nº de unidades secundárias)

X_{ij} - valor da variável em estudo na unidade secundária j do cacho i

$$T_i = \sum_{j=1}^{N_i} X_{ij}$$ - total da variável X no cacho i

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij}$$ - o valor médio de X no cacho i

$N_c = \frac{1}{M}\sum_{i=1}^{M} N_i$ - a dimensão média dos cachos na população

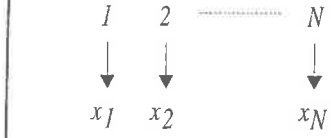
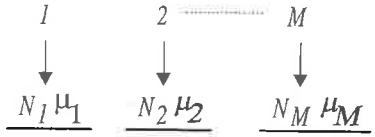
e $\mu = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i} X_{ij}}{N}$ - a média de X na população

onde $N = \sum_{i=1}^{M} N_i$ - representa a dimensão da população

Consideremos aqui o problema de estimação de μ no caso particular de uma selecção aleatória simples de cachos:

$$\mu = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i} X_{ij}}{N} = \frac{\sum_{i=1}^{M} N_i \mu_i}{MN_c} = \frac{1}{M}\sum_{i=1}^{M} \frac{N_i \mu_i}{N_c}$$

Se admitirmos os N_i conhecidos podemos associar a cada unidade "cacho" o valor conhecido $\frac{N_i \mu_i}{N_c}$ que assume um papel equivalente ao valor registado por um determinado indivíduo (relativamente à variável de interesse) num esquema aleatório simples, isto é:

Amostragem aleatória simples	Amostragem em cachos
	
$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\mu = \frac{1}{M}\sum_{i=1}^{M} \frac{N_i\mu_i}{N_c}$

Consequentemente os resultados sobre o estimador $\hat{\mu}$ constituem uma simples aplicação dos resultados apresentados na sondagem aleatória simples.

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} \frac{N_i\mu_i}{N_c}$$

e

$$\text{var}(\hat{\mu}) = \frac{\sigma_c'^2}{m} \cdot \frac{M-m}{M}$$

onde

$$\sigma_c'^2 = \frac{1}{M-1}\sum_{i=1}^{M}\left[\frac{N_i\mu_i}{N_c} - \mu\right]^2$$

A var($\hat{\mu}$) pode então ser estimada por:

$$\widehat{\text{var}}(\hat{\mu}) = \frac{M-m}{Mm} \cdot \frac{\sum_{i=1}^{m}\left[\frac{N_i\mu_i}{N_c} - \hat{\mu}\right]^2}{m-1}$$

Observação: Neste processo a dimensão global da amostra é uma variável aleatória pois depende da amostra de cachos previamente seleccionada.

Sendo N essa v.a. tem-se:

$$N = \sum_{i=1}^{M} N_i \delta_i$$

onde δ_i é uma v.a. de Bernoulli:

$$\delta_i = \begin{cases} 1 & \text{se cacho } i \in \text{ amostra de cachos} \\ 0 & \text{se cacho } i \notin \text{ amostra de cachos} \end{cases}$$

pelo que a dimensão média da amostra vem:

$$E(N) = \sum_{i=1}^{M} N_i E(\delta_i)$$

$$= \sum_{i=1}^{M} N_i \frac{m}{M} = mN_c$$

3) Comparação com a amostragem aleatória simples

Sendo $\hat{\mu}$ o estimador de μ numa amostragem aleatória simples, tem-se:

$$\text{var}(\hat{\mu}) = \frac{N-n}{N} . \frac{\sigma'^2_x}{n}$$

Fazendo $N = MN_C$ e $n = E(N) = mN_C$ resulta

$$\text{var}(\hat{\mu}) = \frac{M-m}{M} . \frac{\sigma'^2_x}{mN_C}$$

Na amostragem em cachos,

$$\text{var}(\hat{\mu}_C) = \frac{M-m}{M} \cdot \frac{\sigma_C'^2}{m}$$

Logo $\text{var}(\hat{\mu}_C) \leq \text{var}(\hat{\mu})$

sse $\sigma_C'^2 \leq \frac{\sigma_x'^2}{N_C}$ ou $\eta^2 = \frac{\sigma_C'^2}{\sigma_x'^2} \leq \frac{1}{N_C}$

onde η^2 representa a percentagem da variância global de X explicada pela variância entre os cachos.

Este resultado suscita a seguinte interpretação:
Para que a sondagem em cachos seja mais precisa que a aleatória simples é necessário que η^2 seja suficiente pequeno (para garantir a condição anterior), isto é, a variância inter-cachos é uma fracção diminuta de σ_x^2 ou seja a existência de **cachos muito homogéneos entre si**.
Complementarmente a dimensão dos cachos não deve ser muito elevado pois nesse caso N_C é elevado o que dificulta a satisfação da condição $\eta^2 \leq 1/N_C$.
Nestes termos, sugere-se a utilização de um maior número de cachos de pequenas dimensões e suficientemente homogéneos.

4.3 - Sondagem em duas etapas

4.3.1 - Introdução

Consideraremos nesta abordagem o recurso à amostragem aleatória simples em cada etapa do processo:

Seja M o número de unidades primárias;

N_i a dimensão da unidade primária *i*

N a dimensão total da população

X_{ij} o valor da variável em estudo no *j-ésimo* indivíduo da unidade primária *i*

Representaremos de novo por *m* a dimensão da "amostra de unidades primárias" e n_i a dimensão da amostra de unidades secundárias provenientes da unidade primária *i*.

Representando por s_i a lista das unidades secundárias na unidade primária *i* inquirida

$$\hat{T}_i = N_i \sum_{j \in s_i} \frac{X_{ij}}{n_i} = N_i \overline{X}_i$$

e $\hat{T}$ pode analogamente definir-se por:

$$\hat{T} = M \sum_{i \in s} \frac{\hat{T}_i}{m}$$

onde *s* representa a lista das unidades primárias inquiridas

isto é,

$$\hat{T} = \frac{M}{m} \sum_{i \in s} \left(\frac{N_i}{n_i} \sum_{j \in s_i} X_{ij} \right)$$

Observação:

1) No caso de uma amostragem aleatória simples o estimador $\hat{T} = N\hat{\mu}$ necessita do conhecimento da dimensão total N da população. Na amostra bietápica não é necessária conhecer N, basta conhecer as dimensões das unidades primárias seleccionadas.

Contudo esta abordagem permite estimar a consideração da estimativa de N:

$$\hat{N} = M \left(\frac{1}{m} \sum_{i \in s} N_i \right)$$

com base no estimador:

$$\hat{N} = M \left(\frac{1}{m} \sum_{i=1}^{M} N_i \delta_i \right)$$

com $\delta_i = \begin{cases} 1 & \text{se } i \in s \\ 0 & \text{se } i \notin s \end{cases}$

Sendo a qualidade desse estimador $\hat{N}$ muito afectada se os N_i das U.P. diferirem consideravelmente entre si.

2) Este processo de amostragem recorre apenas **às bases de sondagem das unidades primárias seleccionadas**

na 1ª etapa, isto é, já não é necessária a base de sondagem associada à população alvo!

3) O estimador linear $\hat{T}$ pode escrever-se na forma:

$$\hat{T} = \sum_{i \in s} \sum_{j \in s_i} \frac{M}{m} \cdot \frac{N_i}{n_i} X_{ij}$$
$$= \sum_{i \in s} \sum_{j \in s_i} W_i X_{ij}$$

o que mostra que a ponderação atribuída ao indivíduo j da unidade primária i é uma característica exclusiva a todas as unidades seleccionadas da unidade primária i. Como, em geral, $\frac{N_i}{n_i}$ varia com i podemos concluir que a amostra bietápica é uma amostragem com probabilidade de inclusão desiguais. Contudo se os n_i forem seleccionados proporcionalmente aos N_i , a amostra bietápica apresentará iguais probabilidades de inclusão.
Este tipo de considerações estendem-se imediatamente ao caso geral de um processo multietápico.

4.3.2 - Estudo da precisão do estimador $\hat{T}$

Para calcular a variância de $\hat{T}$,

$$\hat{T} = M \sum_{i \in s} \frac{\hat{T}_i}{m}$$

vamos recorrer ao seguinte resultado:

$$\operatorname{var}\left(\hat{\mathrm{T}}\right) = \underset{1}{\mathrm{E}}\left(\underset{2|1}{\mathrm{V}}\left(\hat{\mathrm{T}}\right)\right) + \underset{1}{\mathrm{V}}\left(\underset{2|1}{\mathrm{E}}\left(\hat{\mathrm{T}}\right)\right)$$

onde $\underset{1}{\mathrm{V}}$ e $\underset{1}{\mathrm{E}}$ representam, respectivamente a variância e a esperança associadas à selecção das unidades primárias e $\underset{2|1}{\mathrm{V}}$, $\underset{2|1}{\mathrm{E}}$ representam, respectivamente, a variância e a esperança condicionais associadas às extracções das unidades secundárias sabendo que já se procedeu à selecção das unidades primárias.

Obtendo-se:

$$\operatorname{var}\left(\hat{\mathrm{T}}\right) = M^2\left(1-f\right)\frac{\sigma_1'^2}{m} + \frac{M}{m}\sum_{i=1}^{M} N_i^2\left(1-f_i\right)\frac{\sigma_{2,i}'^2}{n_i}$$

onde $$\sigma_1'^2 = \frac{1}{M-1}\sum_{i=1}^{M}\left(\mathrm{T}_i - \mu_p\right)^2$$

com $$\mu_p = \frac{1}{M}\sum_{i=1}^{M}\mathrm{T}_i$$

e $$\sigma_{2,i}'^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left(X_{ij} - \mu_i\right)^2$$

com $$\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}$$

Observação:

1) O primeiro termo da $V(\hat{T})$ está relacionado com a selecção das unidades primárias, exprimindo pois uma **dispersão inter-unidades** primárias avaliada pela dispersão dos totais por unidade primária. Por sua vez o 2º termo reflecte a dispersão da variável *X* nas unidades primárias, isto é, uma **dispersão intra-unidades primárias**.

Consequentemente, para aumentar a precisão do estimador $\hat{T}$ sugere-se:

- uma amostra de unidades primárias de **dimensão *m* elevada**, embora condicionada por critérios decorrentes do orçamento global da sondagem;
- unidades primárias de **dimensão N_i semelhante e pequena**
- **unidades primárias com comportamentos médios semelhantes entre si** de tal forma que a variabilidade das observações seja o mais possível do tipo **"intra-unidades primárias"**.

2) Relativamente a este último comentário trata-se de contrariar a tendência dominante na prática segundo a qual a dispersão dos totais $\sigma_1'^2$ tende a ser elevada, penalizando o 1º termo da $V(\hat{T})$, o mesmo já não se verificando em $\sigma_{2,i}'^2$, o qual tende a exprimir dispersões entre indivíduos directamente comparáveis.

3) A estimativa de $V(\hat{T})$ vem definido por:

$$\widehat{V(\hat{T})} = M^2(1-f)\frac{s_1'^2}{m} + \frac{M}{m}\sum_{i=1}^{m} N_i^2(1-f_i)\frac{s_{2,i}'^2}{n_i}$$

com

$$s_1'^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{T}_i - \frac{\hat{T}}{M}\right)^2$$

ou

$$= \frac{1}{m-1}\left[\sum_{i=1}^{m}\hat{T}_i^2 - \frac{m}{M^2}\hat{T}^2\right]$$

e

$$s_{2,i}'^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{ij} - \hat{\mu}_i\right)^2$$

$$= \frac{1}{n_i - 1}\left[\sum_{j=1}^{n_i}x_{ij}^2 - n_i\left(\hat{\mu}_i\right)^2\right]$$

Nota: Pode-se provar que o 1° termo de $V(\hat{T})$ sobre-estima o 1° termo de $V(\hat{T})$ enquanto que o 2° termo de $\widehat{V(\hat{T})}$ sub-estima o 2° termo de $\widehat{V(\hat{T})}$. Contudo o estimador $V(\hat{T})$ é um estimador centrado de $V(\hat{T})$.

4) No caso particular onde n_i é proporcional a N_i na 2ª etapa do processo, resulta para $\hat{T}$,

$$\hat{T} = \frac{1}{f_1 f_2}\sum_{i\in s}\sum_{j\in s_i}X_{ij}$$

onde $f_1 = \frac{m}{M}$ e $f_2 = \frac{n_i}{N_i}$

e $$V(\hat{T}) = M(1-f_1)\frac{\sigma_1'^2}{f_1} + \frac{1-f_2}{f_1 f_2}\left(\sum_{i=1}^{M} N_i \sigma_{2,i}'^2\right)$$

Se, cumulativamente, todas as unidades primárias tiverem a mesma dimensão $N_i = \frac{N}{M}$ resulta imediatamente, $\hat{T} = N\hat{\mu}$, como era de esperar.

5) O problema da dimensão da amostra

Pretende-se determinar o número m de unidades primárias a considerar na 1ª etapa assim como o número de unidades secundárias n_i a seleccionar em cada uma das unidades primárias i.
O problema pode, na sua forma mais simples reduzir-se ao cálculo de m e de n_i $(i = 1,\cdots,m)$ de forma a proporcionar um erro absoluto ε para as estimativas, fixando um nível de confiança $100(1-\alpha)\%$, isto é:
Sendo $c = \chi_{\alpha/2}$ de uma distribuição $N(0,1)$ tem-se $c\sqrt{\text{var}(\hat{T})} = \varepsilon$

Os valores que figuram nos dois termos da $\text{var}(\hat{T})$ não são em geral conhecidas e terão que ser estimadas por recurso a um eventual conhecimento já existente numa população semelhante ou por via de uma pré-amostra. No caso frequente de trabalharmos com uma amostragem bietápica autoponderada (os indivíduos têm a mesma ponderação sendo fixa a dimensão da amostra), prova-se que o 2° termo da $\text{var}(\hat{T})$ pode ser desprezado o que facilita consideravelmente o processo (P. Ardilly, 1994).

Observação:

Por vezes estabelece-se alguma confusão entre a amostragem multi-etápica e um processo de recolha de informação em **várias fases**, confundindo-os quando na verdade correspondem a abordagens distintas onde convirá desde já precisar que o processo de recolha em várias fases não é propriamente um método de sondagem mas sim uma combinação de várias sondagens sucessivas as quais incidem sobre fracções variáveis da população.
Assim podemos numa **1ª fase**, por recurso a um determinado método de sondagem, construir uma amostra com uma dimensão relativamente elevada inquirindo uma ou poucas características essenciais (o que pode ser muito útil na ausência de uma base de sondagem), e numa **2ª fase, com base na amostra recolhida na 1ª fase**, realizar um inquérito mais aprofundado.

Exemplo:
Pretende-se realizar um inquérito junto da população de trabalhadores Africanos que trabalhem na cidade de Lisboa no sector da construção: não sendo possível dispor de uma lista exaustiva dessas pessoas com os respectivos endereços mas apenas uma listagem dos quarteirões onde habitam essencialmente tais trabalhadores poderíamos **numa 1ª fase** considerar uma amostra aleatória de alojamentos seleccionados em tais quarteirões, isto é, através de uma operação rápida podemos saber em quais alojamentos residem os trabalhadores africanos desse ramo de actividade e então, **numa 2ª fase**, através de um determinado desenho amostral, constituírmos uma amostra reduzida de trabalhadores nessas condições, os quais responderão ao questionário propriamente dito.

CAPÍTULO IV - RECURSO À INFORMAÇÃO AUXILIAR COM VISTA À MELHORIA DA ESTIMAÇÃO

1 - O problema da recomposição da amostra

A disponibilidade de informação auxiliar pode conduzir a uma melhoria no processo de definição de desenho amostral assim como propiciar estimadores mais "precisos" que os estimadores "não informativos" tomados até aqui: em termos práticos quando se dispõe de várias variáveis auxiliares podemos mesmo recorrer a um processo misto onde algumas variáveis são utilizadas para "melhorar o desenho amostral" e outras servem para "melhorar o estimador".
O recurso à informação auxiliar no desenho amostral já foi abordado nos capítulos anteriores pelo que nos circunscreveremos agora ao problema **da melhoria do estimador**: é o chamado problema **da recomposição ou redireccionamento da amostra**.

A filosofia subjacente à "recomposição da amostra" segundo uma dada variável auxiliar X_i resume-se a construir um estimador para o total da variável em estudo capaz de estimar com **variável nula** o total, suposto conhecido, da variável auxiliar.
Suponhamos, a título de exemplo, que seleccionaremos através de uma sondagem aleatória simples uma amostra de 2 000 indivíduos obtendo-se 800 homens e 1 200 mulheres num determinado escalão etário.

O estimador "não informativo" de proporção de homens desse grupo etário existente na população em jogo seria $\hat{p} = \frac{800}{2000} = 40\%$.

No entanto se recorrermos à informação auxiliar proveniente dos dados do recenseamento da população produzido pelo INE, tomaremos conhecimento que nessa população há, por exemplo, 48% de homens e 52% de mulheres evidenciando que o valor $\hat{p}$ é manifestamente errado.

O objectivo de recomposição seria o de construir um estimador $\hat{\theta}$ de um qualquer parâmetro de interesse θ tal que se $\theta = p$ a estimativa de θ seja exactamente p. O que corresponde a alterar as ponderações dos indivíduos na amostra, isto é, a recomposição opera-se nas novas ponderações e não na amostra a qual é, evidentemente, irreversível.

2 - PÓS-ESTRATIFICAÇÃO

2.1 - Introdução

Frequentemente a natureza do problema em estudo sugere a partição da população em estratos, mas não se dispõe das listas de unidades estatísticas correspondentes a cada estrato.
Uma alternativa consiste em proceder a uma amostragem aleatória simples, repartindo posteriormente a amostra segundo os estratos e para efeitos de estimação proceder à recomposição da amostra.
A pós-estratificação é a via mais utilizada para realizar a recomposição da amostra.
Suponhamos que obtivemos por uma amostragem aleatória simples uma amostra com uma dimensão n .
Por intermédio de uma ou mais variáveis auxiliares cujo valor é suposto conhecido para cada indivíduo, constituímos, no seio da amostra, E categorias chamadas pós-estratos as quais correspondem naturalmente a modalidades dessas variáveis auxiliares (ou do seu cruzamento), de tipo qualitativo ou quantitativo.

Seja N_h o efectivo global de cada pós-estrato $h(h = 1,\cdots,H)$ e $\overline{X}_h$ a média da variável X no pós-estrato h.

O estimador $\hat{T}_{post} = \sum_{h=1}^{H} N_h \overline{X}_h$ é obviamente um estimador centrado do parâmetro T, sendo:

$$\hat{\mu}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{X}_h$$

Para isto basta conhecer N_h isto é, o nº de indivíduos da população que pertencem a cada categoria h, isto é, enquanto na estratificação a dimensão da amostra oriunda de cada estrato é previamente estabelecida segundo um dado critério, na pós-estratificação associada a uma sondagem aleatória simples, a dimensão n_h é uma variável aleatória com lei hipergeométrica $H\left(N, n, \frac{N_h}{N}\right)$ e tal que $\sum_{h=1}^{H} n_h = n$ e consequentemente $E(n_h) = \frac{n}{N} N_h$. Aliás o estimador $\hat{T}_{pos}$ pode ser usado para estimar os efectivos reais por cada categoria h com a condição de "colagem" $\sum_{h=1}^{H} n_h = n$. Neste contexto a variável de interesse é uma variável indicatriz que referencia cada indivíduo da população à categoria h à qual ele pertence:

$$\text{Seja } \delta_\alpha^i = \begin{cases} 1 & \text{se } i \in \text{ categoria } \alpha \\ 0 & \text{se } i \notin \text{ categoria } \alpha \end{cases}$$

$$T_\alpha = \sum_{i=1}^{N} \delta_\alpha^i = N_\alpha$$

e

$$\hat{T}_{\alpha_{(post)}} = \sum_{h=1}^{H} N_h \overline{X}_{\alpha,h}$$

$$\text{com } \overline{X}_{\alpha,h} = \frac{\sum_{i \in s_h} \delta_\alpha^i}{n_h} = \begin{cases} 0 & \text{se } \alpha \neq h \\ 1 & \text{se } \alpha = h \end{cases}$$

onde s_h identifica os indivíduos da amostra que pertencem à categoria h, e por conseguinte $\hat{T}_{\alpha_{(post)}} = N_\alpha .1 = N_\alpha$ ou seja o estimador $\hat{T}_{\alpha_{(post)}}$ é pois um estimador com variância nula, isto é, a propriedade fundamental da **recomposição** está verificada.

Observações:

1) Sendo n_h uma variável aleatória a análise da $E\left(\hat{\mu}_{post}\right)$ resulta da consideração inicial da esperança condicional $E\left(\hat{\mu}_{post} / n_h\right)$:

$$E\left(\hat{\mu}_{post} / n_h\right) = \sum_{h=1}^{H} \frac{N_h}{N} E\left(\overline{X}_h / n_h\right)$$

$$= \sum_{h=1}^{H} \frac{N_h}{N} \mu_h = \mu$$

$$\text{logo } E\left(\hat{\mu}_{post}\right) = E\left[E\left(\hat{\mu}_{post} / n_h\right)\right] = \mu$$

2) A pós-estratificação requer informação **à priori** sobre a importância relativa $\frac{N_h}{N}$ de cada estrato na população e informação **à posteriori** permitindo repartir os indivíduos da amostra por cada um dos pós-estratos.

3) Sendo $\hat{\mu}_{post} = \sum_{h=1}^{H} \sum_{j=1}^{n_h} \frac{N_h}{N n_h} X_{hj}$ constata-se que **o "mecanismo de recomposição da amostra"** consiste em modificar a ponderação inicial. Admitindo um processo inicial aleatório simples, a ponderação inicial atribuída a cada indivíduo $i(i = 1, \cdots, n)$ passa de $W_i = \frac{1}{n}$ para $W_i^h = \frac{N_h}{N n_h}$ se o indivíduo $i \in$ à categoria *h*.

4) A pós-estratificação pode adaptar-se ao tratamento de outliers desde que possamos ponderar os pós-estratos da população que contém os indivíduos atípicos.

Exemplo: Pretende-se realizar um inquérito aos trabalhadores da TAP e suponhamos que a natureza do inquérito sugere o interesse em tomar o tipo de adesão sindical de cada trabalhador como critério de estratificação. Contudo, esse tipo de dados individuais são **confidenciais**. Então podemos recorrer a uma pós-estratificação admitindo que podemos obter apenas a percentagem $\frac{N_h}{N}$ de trabalhadores em cada uma das categorias previamente definidas de actividade sindical, bastando então perguntar aos indivíduos da amostra a qual dessas categorias pertencem.

2.2 - Estudo da precisão do estimador $\hat{\mu}_{post}$ na sequência de uma sondagem aleatória simples

O estudo da variância do estimador $\hat{\mu}_{post}$,

$$\hat{\mu}_{post} = \sum_{h=1}^{H} \sum_{j=1}^{n_h} \frac{N_h}{N n_h} X_{hj}$$

complica-se pelo facto de n_h ser aqui uma variante aleatória que figura no denominador da expressão e não existir uma forma de escrever exactamente a esperança matemática do inverso de uma variável aleatória.

Considerando n suficientemente elevado e $\frac{N_h}{N} \geq 0{,}1$ para cada h podemos escrever aproximadamente,

$$E\left(\frac{1}{n_h}\right) \approx \frac{1}{E(n_h)}\left(1 + \frac{\text{var}(n_h)}{\left[E(n_h)\right]^2}\right)$$

e prova-se que

$$\mathbf{var}\left(\hat{\mu}_{post}\right) = \frac{1-f}{n}\left(\sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2\right) + \frac{1-f}{n^2}\left(\sum_{h=1}^{H}\left(1 - \frac{N_h}{N}\right)\sigma_h^2\right)$$

(consultar Gourieroux, pág. 110).

Observações:

1) O 1° termo de $\text{var}\left(\hat{\mu}_{post}\right)$ coincide com a $\text{var}\left(\hat{\mu}_{st}\right)$ **numa amostragem estratificada proporcional** e portanto

$$\text{var}\left(\hat{\mu}_{st}\right) \leq \text{var}\left(\hat{\mu}_{post}\right)$$

o que faz sentido já que a **pós-estratificação é menos exigente em informação auxiliar que a estratificação à priori**: de facto, na pós-estratificação não recorremos à informação auxiliar para cada indivíduo da população, mas apenas ao conhecimento do total N_h de indivíduos que pertencem à categoria h.

2) A convergência do 2° termo para zero é de ordem $\left(\frac{1}{n^2}\right)$ o que significa que para n relativamente grande, este 2° termo é diminuto ou seja para n grande o estimador $\hat{\mu}_{post}$ apresenta uma precisão próxima da obtida numa amostragem estratificada proporcional.

3) A presença de σ_h^2 nos dois termos de $\text{var}\left(\hat{\mu}_{post}\right)$ aconselha a tomar pós-estratos assimilados a sub-populações bem homogéneas, isto é, alicerçados em variáveis auxiliares com um forte poder explicativo da variável em estudo.

4) Supondo para cada h, $n_h \geq 2$

$$\widehat{\text{var}\left(\hat{\mu}_{post}\right)} = \frac{1-f}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}\hat{\sigma}_h^2\right) + \frac{1-f}{n}\left(\sum_{h=1}^{H}\left(1-\frac{N_h}{N}\right)\hat{\sigma}_h^2\right)$$

Com $\hat{\sigma}_h^2 = S_h'^2$

Nota: a condição $n_h \geq 2$ para todo o h pode não ser satisfeita se tormamos um nº elevado H de categorias ou se n for pequeno.

5) A pertinência da pós-estratificação depende particularmente de fiabilidade da informação relativa a N_h .

Suponhamos então que o valor conhecido não é o valor exacto mas uma quantidade aproximada N_{0h}/N_0 :

então o estimador $(\hat{\mu})_{post}$ vem

dado por $(\hat{\mu}_0)_{post} = \sum_{h=1}^{H} \frac{N_{oh}}{N_0} . \overline{X}_h$

em vez de $(\hat{\mu})_{post} = \sum_{h=1}^{H} \frac{N_h}{N} . \overline{X}_h$

$$\Rightarrow (\hat{\mu}_0)_{post} - (\hat{\mu})_{post} = \sum_{h=1}^{H} \left(\frac{N_{oh}}{N_o} - \frac{N_h}{N} \right) \overline{X}_h$$

$$E(\hat{\mu}_0)_{post} = \mu + \underbrace{\sum_{h=1}^{H} \left(\frac{N_{oh}}{N_o} - \frac{N_h}{N} \right) E(\overline{X}_h)}$$

enviesamento

Este resultado evidência não só o enviesamento na sequência de termos tomado $\frac{N_{0h}}{N_0}$ e não o valor exacto $\frac{N_h}{N}$ mas também o facto preocupante desse **enviesamento não depender da dimensão da amostra.**

6) No caso de recorrermos a várias variáveis auxiliares na definição dos critérios de pós-estratificação e admitindo conhecida a estrutura da população para cada cruzamento desses critérios, o problema reduz-se naturalmente ao caso anterior por recurso ao estimador $\hat{\mu}_{post}$.

Contudo, é frequente não dispormos da estrutura da população mediante o cruzamento de critérios, mas apenas acedermos aos totais por categoria relativamente a cada uma das variáveis tomada isoladamente. O algoritmo do "Raking Ratio" (consultar P.Ardilly, pág. 216) fornece uma solução de "colagem" para este tipo de problemas.

Exemplo (Droesbeke J., Lavallé P. 1998)

Pretende-se estimar o rendimento total da população de 100 proprietários de um quarteirão residencial. Sabe-se que a distribuição desses proprietários por sexo é tal que 60% são homens e 40% são mulheres. Para o efeito, suponhamos que se toma uma amostra de 8 proprietários quaisquer escolhidos na população, registando-se o rendimento e o sexo do proprietário,

Proprietário i	Rendimento x_i (10^3 Esc.)	Sexo
1	450	M
2	300	M
3	360	M
4	420	M
5	450	F
6	660	F
7	420	F
8	510	F

a partir da amostra a estimativa do rendimento total será

$$\hat{T}_x = N\overline{X} = 100\left(\frac{\sum_{i=1}^{8} X_i}{8}\right)$$

$$= 12.5\left(\sum_{i=1}^{8} X_i\right) = 44625$$

Contudo esta estimativa pode ser melhorada fazendo apelo à informação auxiliar sobre a repartição do sexo na população em estudo, isto é, faz sentido recorrer ao estimador pós-estratificado:

$$\left(\hat{T}_x\right)_{post} = \sum_{h=1}^{2} N_i \overline{X}_h$$

$$= \sum_{i=1}^{2} \frac{N_i}{n_h}\left(\sum_{j=1}^{n_h} X_{hj}\right)$$

$$= \frac{60}{4}\left(\sum_{j=1}^{4} X_{1j}\right) + \frac{40}{4}\left(\sum_{j=1}^{4} X_{2j}\right)$$

$$= 15(1530) + 10(2040)$$

$$= 43350$$

3 - ESTUDO DO ESTIMADOR DO QUOCIENTE

3.1 - Introdução

A razão entre duas médias amostrais $\bar{x}/\bar{y}$ pode ser utilizada, sob certas condições, como estimativa $\hat{R}$ de verdadeira razão R entre μ_x e μ_y. O conhecimento dessa estimativa permite estimar o total, a média (ou população) da variável de interesse X conhecendo o comportamento na população de variável auxiliar Y:

$$\hat{T}_x = \hat{R}\, T_y \qquad \hat{\mu}_x = \hat{R}\, \mu_y \qquad \text{com } \hat{R} = \frac{\overline{X}}{\overline{Y}}$$

isto é, o objectivo deste método é melhorar a precisão do estimador de variável de interesse recorrendo à informação auxiliar fornecida por outra variável disponível na população; mostraremos que essa variável auxiliar deverá apresentar **uma elevada correlação linear** com a variável em estudo.

Exemplo: Considerando que a variável X, representando a população residente nas secções estatísticas do País em 2001, apresenta uma elevada correlação linear com a variável Y representando a população residente nas secções estatísticas no último recenseamento da população (1991), poderemos em poucas semanas após o início da operação censitária 2001 obter a população residente numa amostra de secções estatísticas, definidas previamente mediante um conveniente desenho amostral, estimando assim a evolução R da população residente entre 1991 e 2001 e consequentemente fornecendo uma primeira estimativa da população residente em 2001.

3.2 - Estudo dos estimadores $\hat{R}, \hat{T}_X$ e $\hat{\mu}_x$

3.2.1 - $\hat{R}$ não é um estimador centrado do parâmetro $R = \mu_x / \mu_y$

com efeito

$$\begin{aligned} \operatorname{cov}(\hat{R},\hat{Y}) &= E(\hat{R}.\hat{Y}) - E(\hat{R})E(\hat{Y}) \\ &= \mu_x - \mu_y E(\hat{R}) \end{aligned}$$

$$\Rightarrow E(\hat{R}) = R - \frac{1}{\mu_y}\operatorname{cov}(\hat{R},\hat{Y})$$

isto é, o estimador $\hat{R}$ apresenta um enviesamento

$$E(\hat{R}) - R = -\frac{1}{\mu_y}\operatorname{cov}(\hat{R},\overline{Y})$$

Vamos estudar a parte principal desse enviesamento e mostrar que $\hat{R}$ é assintoticamente centrado:

$$\hat{R} - R = \frac{\overline{X} - R\overline{Y}}{\overline{Y}}$$

escrevendo $\dfrac{1}{\overline{Y}} = \dfrac{1}{\mu_y}\left(1 + \dfrac{\overline{Y} - \mu_y}{\mu_y}\right)^{-1}$

e desenvolvendo em série na vizinhança de zero, tem-se:

$$\hat{R} - R = \frac{\overline{X} - R\overline{Y}}{\mu_y}\left[1 - \frac{\overline{Y} - \mu_y}{\mu_y} + \frac{\left(\overline{Y} - \mu_y\right)^2}{\mu_y{}^2} - \cdots\right]$$

tomando a aproximação fornecida pelos dois primeiros termos da série:

$$\hat{R} - R \approx \frac{\overline{X} - R\overline{Y}}{\mu_y}\left(1 - \frac{\overline{Y} - \mu_y}{\mu_y}\right)$$

donde

$$E\left(\hat{R} - R\right) \approx -\frac{1}{\mu_y^2}\left[\operatorname{cov}\left(\overline{X}, \overline{Y}\right) - R\operatorname{var}\left(\overline{Y}\right)\right]$$

pelo que o enviesamento de $\hat{R}$ é aproximadamente definido por

$$E\left(\hat{R}\right) - R \approx \frac{1-f}{n\mu_y^2}\left[R\sigma_y^2 - \rho\sigma_x\sigma_y\right] \quad \begin{matrix} \to 0 \\ n \to \infty \end{matrix}$$

onde ρ representa o coeficiente de correlação linear entre X e Y

O enviesamento pode ainda escrever-se em termos dos coeficientes de variação de X e Y:

$$E\left(\hat{R} - R\right) \approx \frac{1-f}{n} R\left[\left[cv(Y)\right]^2 - \rho cv(Y).cv(X)\right]$$

Observação: O enviesamento é nulo sse

$$R\sigma_y^2 - \rho\sigma_x\sigma_y = 0$$

isto é, $\dfrac{\rho\sigma_x\sigma_y}{\sigma_y^2} = \dfrac{\sigma_{xy}}{\sigma_y^2} = \dfrac{\mu_y}{\mu_x}$

Condição satisfeita, se a recta de regressão de X sobre Y passar pela origem (representando σ_{xy} / σ_y^2 o declive dessa recta).

3.2.2 - Estudo de precisão do estimador $\hat{R}$ para *n* grande

$$\text{var}\left(\hat{R}\right) \approx E\left(\hat{R} - R\right)^2 \approx \frac{1}{\mu_y^2} E\left(\overline{X} - R\overline{Y}\right)^2$$

isto é, $\text{var}\left(\hat{R}\right) \approx \dfrac{1}{\mu_y^2}\text{var}\left(\overline{\varepsilon}\right)$

onde $\overline{\varepsilon} = \dfrac{1}{n}\sum_{i=1}^{n}\varepsilon_i$ e $\varepsilon_i = X_i - RY_i$

$$\Rightarrow \text{var}\left(\hat{R}\right) \approx \frac{1}{\mu_y^2}.\frac{\sigma_\varepsilon^2}{n}.\frac{N-n}{N-1}$$

resultando

$$\mathbf{var}\left(\hat{R}\right) \approx \frac{1-f}{n\mu_y^2}.\frac{\sum_{i=1}^{N} X_i - RY_i}{N-1}$$

Observações:

1)

$$\text{var}\left(\hat{T}_x\right) = \left(T_y\right)^2 \text{var}\left(\hat{R}\right)$$

$$\text{e} \quad \text{var}\left(\hat{\mu}_x\right) = \left(\mu_y\right)^2 \text{var}\left(\hat{R}\right)$$

2) Sendo $\hat{\mu}_x$ um estimador enviesado a estimação pelo quociente não produz em geral bons resultados para n de pequena dimensão e o erro quadrático médio pode ser considerável. Para n suficientemente elevado $(n \geq 30)$ é então razoável assimilar esse erro à variância.

3) Se considerarmos a variável de interesse X relacionada com a variável auxiliar Y segundo uma relação aproximadamente linear, isto é, segundo um modelo da forma:

$$X_i = RY_i + \varepsilon_i$$

onde ε_i é termo residual tal que $\sum_{i=1}^{N} \varepsilon_i = 0$ se $R = \dfrac{\mu_x}{\mu_y}$

então

$$\text{var}\left(\hat{\mu}_x\right) \approx \left(1-f\right) \frac{\sum_{i=1}^{N} \left(X_i - RY_i\right)^2 / (N-1)}{n}$$

$$\approx \left(1-f\right) \frac{\sigma'^2_\varepsilon}{n} \qquad (n \geq 30)$$

o que mostra que a precisão de $\hat{\mu}_x$ no estimador do quociente tem uma forma semelhante à precisão desse estimador na amostragem aleatória simples: $\sigma_x'^2$ é agora substituído por $\sigma_\varepsilon'^2$, isto é,

$$V(\hat{\mu}_x)_{quoc} < V(\hat{\mu}_x)_{a.a.s.}$$

sse $$\sigma_\varepsilon'^2 < \sigma_x'^2$$

ou $$\sigma_x^2 + R^2\sigma_y^2 - 2R\rho\sigma_x\sigma_y < \sigma_x^2$$

isto é, $$\rho > \frac{R}{2}.\frac{\sigma_y}{\sigma_x} = \frac{1}{2}\left(\frac{cv(Y)}{cv(x)}\right)$$

4) A variância do estimador $\hat{R}$ é estimada por:

$$\widehat{\text{var}}(\hat{R}) = \frac{1-f}{n(\bar{Y})^2}\frac{\sum_{i=1}^{n}(X_i - \hat{R}Y_i)^2}{n-1}$$

onde $$\sum_{i=1}^{n}(X_i - \hat{R}Y_i) = \sum_{i=1}^{n}X_i^2 + \hat{R}^2\sum_{i=1}^{n}Y_i^2 - 2\hat{R}\sum_{i=1}^{n}X_iY_i$$

5) Por vezes dispomos de várias variáveis auxiliares concorrentes: é possível programar um quociente para cada variável auxiliar constituindo em seguida uma combinação linear ponderada de todos os quocientes. Cochran desenvolveu métodos optimais para a selecção dos ponderadores.

Exemplo:

O quadro seguinte mostra o número de habitantes (em milhares) de 49 cidades seleccionadas aleatoriamente dum conjunto de 196. Pretende estimar-se o número total de habitantes destas cidades em 1980. Em 1970 este valor era de 22919.

O exemplo é adequado para o estimador do quociente. A maioria das cidades na amostra aumentou a sua dimensão, em cerca de 20%, de 1970 para 1980. A partir da amostra tem-se,

$$\sum_{i=1}^{49} x_i = 6262 \; ; \qquad \sum_{i=1}^{49} y_i = 5054$$

Consequentemente, o estimador do quociente para o total das 196 cidades, referente a 1980 é,

$$\hat{T}_x = \frac{\overline{x}}{\overline{y}} T_y = \frac{6262}{5054}(22919) = 28397$$

O estimador correspondente, calculado a partir da população média de cada uma das 49 cidades é,

$$\hat{T}_x = N\overline{X} = \frac{196(6262)}{49} = 25048$$

A população exacta de 1980, nas 196 cidades, era de 29351 habitantes.

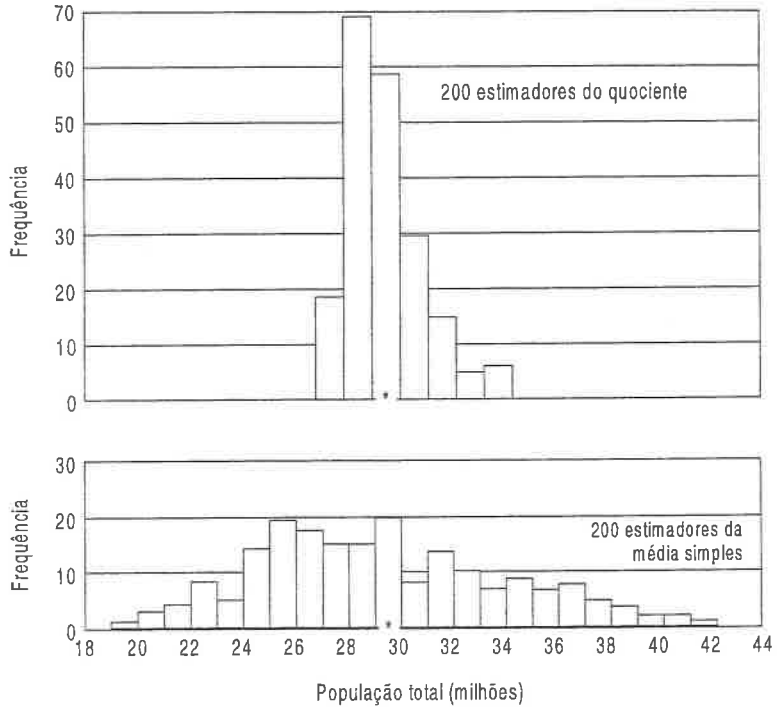
Dimensão das 49 cidades (em milhares de habitantes) em 1970 (y_i) e 1980 (x_i)

y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i
76	80	78	106	64	77	256	288
138	143	66	86	56	142	43	61
67	67	60	57	40	60	25	57
29	50	46	65	40	64	94	85
381	464	2	50	38	52	43	50
23	48	507	634	136	139	298	317
37	63	179	260	116	130	36	46
120	115	121	113	46	53	161	232
61	69	50	64	243	291	74	93
387	459	44	58	87	105	45	53
93	104	77	89	30	111	36	54
172	183	64	63	71	79	48	75
				50	58		

A figura seguinte mostra a distribuição de frequência dos valores dos estimadores do quociente e dos estimadores baseados na média por cidade, obtidos em 200 amostras aleatórias de dimensão 49, retiradas do conjunto das 196 cidades.

É notória a melhoria substancial de precisão obtida pelo método de estimação do quociente. Verifica-se uma maior concentração das estimativas em torno do parâmetro populacional que se pretende estimar, traduzida num menor desvio padrão.

Figura: Comparação experimental do estimador do quociente com o estimador baseado na média da amostra.



** assinala o valor correcto da população de 1970.*

4 - ESTIMADOR DO QUOCIENTE NUMA AMOSTRAGEM ESTRATIFICADA

Podemos considerar duas abordagens: na primeira construímos o estimador da média μ_x pelo quociente a partir dos estimadores do quociente elaborados estrato a estrato, ponderados pela importância relativa de cada estrato:

$$\hat{\mu}_x = \sum_{i=1}^{k} \frac{N_i}{N} \left(\mu_{yi} \frac{\overline{X}_i}{\overline{Y}_i} \right)$$

chama-se **estimador separado do quociente** e a sua utilização supõe o conhecimento da média da variável auxiliar em cada das sub-populações que constituem os estratos.

Alternativamente, em vez de tomarmos uma média ponderada de estimadores pelo quociente, estabelece-se um rácio de médias:

$$\hat{\mu}_x = \mu_y \cdot \frac{\sum_{i=1}^{k} \frac{N_i}{N} \overline{X}_i}{\sum_{i=1}^{k} \frac{N_i}{N} \overline{Y}_i}$$

isto é, o quociente entre as médias estimadas segundo a estratificação para X e para Y, multiplicando pelo verdadeiro valor da média da variável auxiliar Y que é suposta conhecida.

Este estimador, designado **estimador combinado do quociente** não requer o conhecimento da média da variável auxiliar em cada estrato, mas apenas a média dessa variável na população global, mas supondo que o quociente permanece relativamente estável de um estrato para o outro.

Observações:

1) O estimador separado do quociente pode escrever-se na forma:

$$\hat{T}_{x(sep)} = \sum_{i=1}^{K} \hat{R}_i T_{yi}$$

e $$\widehat{\text{var}\left(\hat{T}_x\right)}_{(sep)} = \sum_{i=1}^{k} \left(\hat{T}_{yi}\right)^2 \widehat{\text{var}\left(\hat{R}_i\right)}$$

2) O estimador combinado do quociente pode assumir a forma:

$$\left(\hat{T}_x\right)_c = \frac{\hat{\mu}_x}{\hat{\mu}_y} . T_y = \hat{R} T_y$$

provando-se facilmente que:

$$\text{var}\left(\hat{T}_x\right)_c = \sum_{i=1}^{k} \frac{N_i^2 (1 - f_i)}{n_i} \left(\sigma_{xi}^2 + R^2 \sigma_{yi}^2 - 2R\sigma_{xyi}\right)$$

e por conseguinte:

$$\widehat{\text{var}\left(\hat{T}_x\right)}_c = \sum_{i=1}^{k} \frac{N_i^2 (1 - f_i)}{n_i} \left(\hat{\sigma}_{xi}^2 + \hat{R}^2 \sigma_{yi}^2 - 2\hat{R}\hat{\sigma}_{xyi}\right)$$

3) Os estimadores $\left(\hat{T}_x\right)_{sep}$ e $\left(\hat{T}_x\right)_{comb}$ são estimadores concorrentes.

A selecção do estimador a tomar baseia-se em 2 tipos essenciais de considerações:

- A variabilidade de R_i de estrato para estrato

- A dimensão amostral em cada estrato

Se R_i apresentar uma certa variabilidade de estrato para estrato é preferível o estimador separado.
Complementarmente prova-se que qualquer que seja ***n***, a $\text{var}\left(\hat{T}_x\right)_{sep} < \text{var}\left(\hat{T}_x\right)_{comb}$. Logo quando se trabalha com grandes amostras (≥ 30) em todos os estratos o estimador separado é preferível.

Mas se em cada estrato a amostra não for suficientemente elevada para admitir a aproximação de variância do estimador é conveniente considerar o estimador combinado.
Esse estimador também se recomenda se o quociente R_i variar pouco de estrato para estrato.

4.1 - O estimador do quociente no contexto geral de uma amostragem probabilística com probabilidades desiguais.

Podemos definir $\hat{\mu}_x = \frac{\overline{X}}{\overline{Y}}\mu_y$ onde $\overline{X}$ e $\overline{Y}$ são os estimadores de Horvitz-Thompson de μ_x e μ_y respectivamente:

$$\overline{X} = \frac{1}{N}\sum_{i\in s}\frac{X_i}{p_i} \quad \text{e} \quad \overline{Y} = \frac{1}{N}\sum_{i\in s}\frac{X_i}{p_i}$$

Neste caso, os resultados principais desenvolvidos no contexto de uma sondagem aleatória simples mantém as mesmas características, a saber:

- O estimador $\hat{\mu}_x$ é enviesado com um cruzamento que tende para zero com $\frac{1}{n}$

- Variância de $\hat{\mu}_x$ é aproximadamente igual à variância do estimador centrado da média dos resíduos:

$$v(\hat{\mu}_x) = v\left(\frac{1}{N}\sum_{i\in s}\frac{\varepsilon_i}{p_i}\right)$$

onde $\varepsilon_i = Y_i - RX_i$

A precisão do estimador $\hat{\mu}_x$ depende então da ordem de grandeza dos resíduos ε_i.

5 - ESTIMADOR DE REGRESSÃO LINEAR

Suponhamos que a variável de interesse X e a variável auxiliar Y podem ser relacionadas segundo um modelo linear,

$$X_i = a + bY_i + \varepsilon_i$$

portanto mais geral que o modelo implícito no estimador do quociente e com o mesmo objectivo de melhorar a precisão na estimação de um parâmetro relativo a X com base na variável auxiliar Y, cujo média μ_y é suposto conhecido entre as possíveis escolhas de ***a*** e ***b*** que satisfazem a condição:

$$X_i = a + bY_i + \varepsilon_i$$

e tais que $\sum_{i=1}^{N} \varepsilon_i = 0$

a escolha óptima, no sentido dos mínimos quadrados consiste em tomar para verdadeiros valores de ***a*** e ***b***,

$$b = \frac{\sum_{i=1}^{N} (X_i - \mu_x)(Y_i - \mu_y)}{\sum_{i=1}^{N} (Y_i - \mu_y)^2}$$

e $$a = \overline{X} - b\overline{Y}$$

logo podemos escrever

$$\sum_{i=1}^{N} x_i = Na + b\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \varepsilon_i$$

ou $$\mu_x = a + b\mu_y$$

Por outro lado, admitindo que ao nível do aumento da dimensão n a soma dos resíduos é ainda nula tem-se

$$\sum_{i=1}^{n} X_i \approx na + b\sum_{i=1}^{n} Y_i \quad \text{ou} \quad \overline{X} \approx a + b\overline{Y}$$

Logo $$\mu_x - \overline{X} \approx b\left(\mu_y - \overline{Y}\right)$$

ou $$\mu_x \approx \overline{X} + b\left(\mu_y - \overline{Y}\right)$$

logo o estimador de regressão vem definido por:

$$\left(\hat{\mu}_x\right)_{reg} = \overline{X} + \hat{b}\left(\mu_y - \overline{Y}\right)$$

onde $$\hat{b} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

Observações:

1) Prova-se que $\hat{b}$ é um estimador ligeiramente enviesado do parâmetro b **tal como foi definido atrás** e por conseguinte recomenda-se uma amostra relativamente elevada ($n \geq 50$, segundo algumas simulações realizadas) para utilizar o estimador de regressão.

2) Se a verdadeira recta passa pela origem pode provar-se que $(\hat{\mu}_x)_{reg}$ coincide com o estimador de μ_x pelo quociente.

3) O estimador $(\hat{\mu}_x)_{reg}$ é um estimador linear, isto é, de forma

$$\sum_{i \in s} W_i . X_i$$

Sendo $(\hat{\mu}_x)_{reg} = \overline{X} + \hat{b}(\mu_y - \overline{Y})$

e $\hat{b} = \sum_{i=1}^{n} \frac{(Y_i - \overline{Y})}{\sum (Y_i - \overline{Y})^2} X_i$

$$\Rightarrow (\hat{\mu}_x)_{reg} = \sum_{i=1}^{n} \left[\frac{1}{n} + (\mu_y - \overline{Y}) \frac{Y_i - \overline{Y}}{\sum (Y_i - \overline{Y})^2} \right] .$$

isto é, as ponderações W_i dependem do indivíduo i e da amostra s.

4) Por análise com o caso particular do quociente, para n grande e no contexto de uma sondagem aleatória simples,

$$v(\hat{\mu}_x)_{reg} \approx \frac{1-f}{n} . \sigma'^2_{\varepsilon}$$

onde σ'^2_{ε} representa a variância do resíduo do modelo, $\varepsilon_i = X_i - a - bY_i$,

isto é, $\sigma_{\varepsilon}'^2 = \frac{1}{N-1}\sum_{i=1}^{N}\varepsilon_i^2$

ora

$$\sigma_{\varepsilon}^2 = \sigma_x^2 + b^2\sigma_y^2 - 2b\sigma_{xy}$$

$$= \sigma_x^2 + \left(\frac{\sigma'_{xy}}{\sigma_y'^2}\right)^2 \sigma_y^2 - 2\frac{\sigma_{xy}}{\sigma_y^2}\sigma_{xy}$$

$$\sigma_{\varepsilon}^2 = \sigma_x^2 - \frac{(\sigma_{xy})^2}{\sigma_y^2} = \sigma_x^2 - \rho^2\sigma_x^2$$

$$\Rightarrow \sigma_{\varepsilon}'^2 = (1-\rho^2)\sigma_x'^2$$

$$\Rightarrow v(\hat{\mu}_x)_{reg} \approx (1-\rho^2) \quad v(\hat{\mu}_x)_{aas} \leq v(\hat{\mu}_x)_a$$

Conclusão: O estimador $(\hat{\mu}_x)reg$ é preferível a $(\hat{\mu}_x)a.a.s.$ em todos desde que a dimensão da amostra seja elevada sendo o efeito de sondagem resultante de tomar $(\hat{\mu}_x)reg$ em vez de $(\hat{\mu}_x)a.a.s.$ igual a

$$\boldsymbol{Deff = 1-\rho^2}$$

ilustrando bem que a melhoria de precisão conseguida com o estimador da regressão é tanto mais importante quando mais $|\rho|$ é próximo de 1.

5) A variância de $(\hat{\mu}_x)reg$ é estimada por

$$v(\hat{\mu}_x)_{reg} = \frac{1-f}{n}\cdot\frac{1}{n-1}\sum_{i\in s}\left(Y_i - \hat{a} - \hat{b}X_i\right)^2$$

6) Se dispusermos de várias variáveis auxiliares $Y_1, Y_2, \cdots . Y_k$ e admitindo que a variável em estudo é bem explicada por uma combinação linear dessas variáveis podemos considerar o modelo linear geral

$$X_i = b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki} + \mu_i$$

e obter uma generalização do estimador da regressão (Gorieroux, ...)

7) No caso de estimadores da regressão no contexto de planos de sondagem complexos, esses estimadores apresentam em geral uma variância muito elaborada e função do plano de sondagem em consideração (consultar P.Ardilly, 1994).

8) Recentemente J. Deville e C Sarndall (JASA, 418, 1992) desenvolveram uma abordagem geral que inclui todos os métodos estudados de recomposição da amostra.

CAPÍTULO V - ESTIMAÇÃO EM DOMÍNIOS

5.1 - Introdução

Designa-se por domínio de estudo uma sub-população U_d de uma população U, para a qual se pretende estimar determinados parâmetros.

A necessidade de produção de estimativas fiáveis para parâmetros de variáveis de interesse em pequenos domínios (e em particular pequenas áreas) tem vindo a manifestar-se nas mais variadas áreas.

Também no âmbito da sondagens de opinião e em particular das sondagens eleitorais, este problema se coloca com grande acuidade. De facto, é habitual ver-se a apresentação de estimativas referentes a sub-populações definidas segundo critérios diversos. Alguns desses critérios prendem-se com a segmentação regional da população, ou mais simplesmente com a consideração de grupos definidos por variáveis sócio-económicas, como sejam o sexo, a classe etária ou o grupo socio-económico.

Se as unidades de U_d podem ser identificadas *a priori*, então o plano de sondagem poderá contemplar uma dimensão amostral que permita a estimação nesse domínio com precisão adequada. Neste caso, o domínio pode ser entendido como um estrato ou grupo de estratos e é normalmente designado por domínio planeado. No entanto, mesmo quando a identificação das unidades de U_d é possível, a consideração de uma dimensão amostral adequada poderá ser impraticável por razões logísticas ou orçamentais. Tal facto é particularmente provável quando existe interesse sobre um

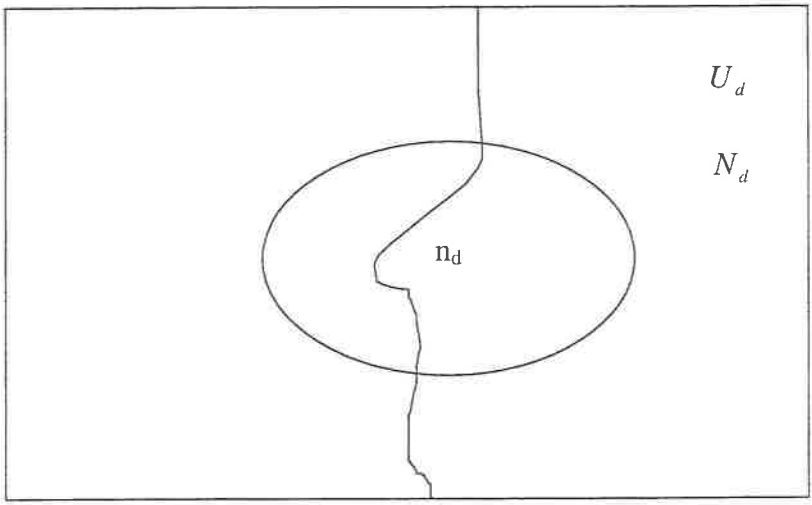
grande número de domínios de pequena dimensão. Neste último caso o domínio diz-se não planeado. São sobretudo estes que criam dificuldades na estimação e que serão abordados neste curso.

Na produção de tais estimativas, emerge um problema evidente, sobretudo em domínios de pequena dimensão, já que as dimensões amostrais no interior de tais domínios têm natureza aleatória e são muitas das vezes demasiado pequenas para que se possam estimativas de precisão aceitável.

Como resultado destas preocupações recentes, têm vindo a ser propostos vários tipos de estimadores que combinem informação auxiliar com os dados da sondagem para os domínios de estudo.

5.2 - Formalização

Recorde-se que um domínio é uma sub-população U_d de uma população U, para a qual se pretende estimar determinados parâmetros. A dimensão do domínio d é representada por N_d.



Representa-se, como habitualmente, a média da variável de interesse na população por μ e a média no domínio por μ_d.

Tem-se então

$$\mu_d = \frac{1}{N_d} \sum_{i \in U_d} y_i .$$

Tendo uma amostra *s*, de dimensão *n*, seleccionada a partir do da população U através de um determinado plano de sondagem, define-se:

$$s_d = U_d \cap s$$
$$n_d = \#(s_d)$$

O problema em questão é o de estimar as características de sub-populações definidas segundo um critério não utilizado numa estratificação *a priori*.

O número de unidades da amostra filtradas no domínio de estudo é aleatório, o que deverá ser tido em conta quando se extrapolam os resultados da amostra para o conjunto da população que constitui o domínio de estudo.

De facto, a aleatoriedade pode ser decomposta em dois níveis:

- Aleatoriedade em n_d.
- Para n_d fixo, há a aleatoriedade da tiragem de n_d unidades de entre as N_d que constituem o domínio.

5.3 - Estimação

Como é bem sabido o objectivo de uma sondagem eleitoral e em geral de qualquer inquérito por amostragem é o de fazer inferência sobre parâmetros de uma população finita.

Os estimadores usados são, muitas vezes, classificados em duas categorias designadas por *design based* e *model based*[1].

[1] Numa abordagem *design based* ou *model-free*, considera-se que as características da população, sejam, no caso univariado, $\{Y_1, Y_2, ..., Y_N\}$, são fixas e que a componente probabilística é introduzida pelo analista ao adoptar determinado plano de amostragem. Na inferência *model based* as características $\{Y_1, Y_2, ..., Y_N\}$ são variáveis aleatórias com valores gerados por uma superpopulação. O modelo poderá ser baseado no conhecimento do fenómeno natural que influencia a distribuição da população ou num modelo estatístico que sumarize as características

Embora, no âmbito da estimação em domínios de pequena dimensão a inferência *model-based* ganhe especial relevo dada a exiguidade da informação disponível, este curso abordará apenas a inferência do tipo *design-based.* Referências mais detalhadas podem ser encontradas em Coelho (1996).

Em particular, e tendo em conta o normal enquadramento das sondagens eleitorais, cingir-nos-emos à análise de estimadores que não utilizam qualquer informação auxiliar para além das dimensões das populações, sob dois planos de sondagem de grande popularidade: a sondagem aleatória simples e a sondagem aleatória estratificada.

5.3.1 - Sondagem aleatória simples

Para estudar as propriedades do estimador de μ_d torna-se útil definir uma variável indicatriz $1_{i\in d}$ que assume o valor 1 se o elemento i da população pertencer ao domínio d e zero caso contrário

Define-se ainda a variável $Z_i = y_i 1_{i\in d}$. Trata-se de uma variável que para o elemento i da população assume valor igual ao da variável de interesse se este pertencer ao domínio e zero caso contrário.

Para tornar a notação mais familiar passar-se-á a representar a variável indicatriz por $X_i = 1_{i\in D}$.

---

fundamentais dessa população. Uma referência mais detalhada sobre este assunto pode ser encontrada em Särndal, Swensson e Wretman (1992).

Vem evidentemente que $y_i = \frac{Z_i}{X_i}$ e consequentemente a média no domínio

$$\mu_d = \frac{1}{N_d}\sum_{i\in d} y_i = \frac{\sum_{i=1}^{N} y_i 1_{i\in d}}{\sum_{i=1}^{N} 1_{i\in d}} = \frac{\tau_z}{\tau_x} = \frac{\mu_z}{\mu_x}.$$

Esta pode ser encarada como a razão entre duas variáveis, pelo que a sua estimação se reduz ao caso da estimação de um quociente.

Vem então

$$\hat{\mu}_d = \bar{Y}_d = \frac{\bar{Z}}{\bar{X}} = \frac{\sum_{i\in s} y_i 1_{i\in d}}{\sum_{i\in s} 1_{i\in d}} = \frac{1}{n_d}\sum_{i\in s_d} y_i .$$

No entanto, demonstra-se (ver Grosbras, 1987), que apesar de se tratar da estimação de um quociente o enviesamento do estimador é praticamente nulo, pelo que se poderá considerar que $\bar{Y}_d$ é um estimador aproximadamente centrado de μ_d.

O erro quadrático médio do estimador da média no domínio vem assim aproximadamente igual à sua variância

$$V_1(\bar{Y}_d) \approx EQM(\bar{Y}_d) \approx \frac{N-n}{N_d n}\sigma'^2_{yd} + \frac{(N-n)(N-N_d)}{(N_d n)^2}\sigma'^2_{yd}.$$

Para domínios de dimensão razoável o 2º termo da expressão pode ser negligenciado vindo numa aproximação mais simplificada

$$V_2(\bar{Y}_d) \approx EQM(\bar{Y}_d) \approx \frac{N-n}{N_d n}\sigma'^2_{yd}$$

A 2ª aproximação coincide com a variância exacta do estimador da média μ_d que resultaria da extracção de uma amostra de dimensão de dimensão $n_d^0 = E(n_d)$ a partir das N_d unidades que constituem a população U_d.

Como tal, a comparação entre as duas aproximações $V_1(\bar{Y}_d)$ e $V_2(\bar{Y}_d)$ indica a perda de precisão resultante da aleatoriedade da dimensão amostral n_d, i.e., da incapacidade de identificar antecipadamente o domínio.

A razão entre as duas aproximações é:

$$\frac{V_1(\bar{Y}_d)}{V_2(\bar{Y}_d)} = 1 + \frac{N-N_d}{N_d n} = 1 + \frac{N-N_d}{N n_d^0} \approx 1 + \frac{1}{n_d^0}$$

Repare-se que quando a dimensão amostral esperada no domínio $n_d^0 = E(n_d)$ é razoavelmente grande, a perda de precisão é negligenciável. No entanto, para dimensões amostrais muito pequenas poderá haver um acréscimo de variância apreciável.

Para encontrar um estimador da variância, no caso em que a dimensão do domínio é desconhecida, será necessário ter em conta que N_d é uma quantidade que terá igualmente que ser estimada,

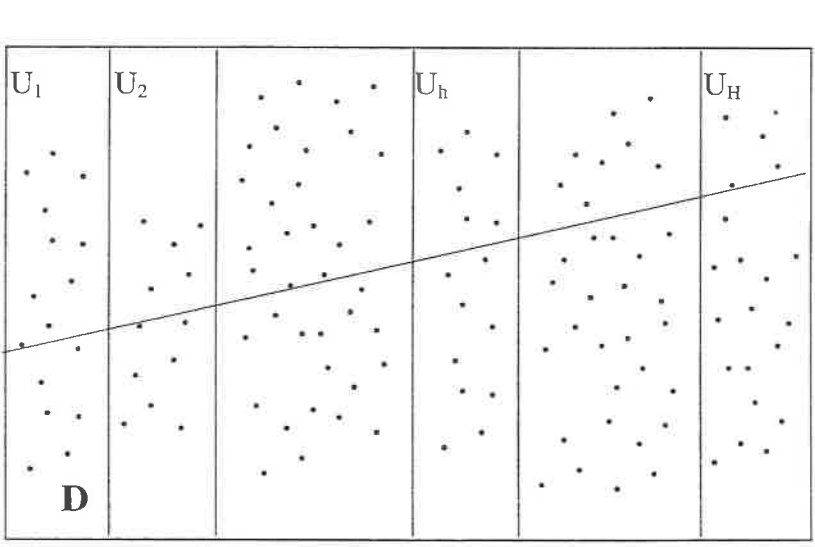
$$\hat{N}_d = \frac{n_d}{n} N .$$

O estimador da variância vem então

$$\hat{V}(\overline{Y}_d) = \frac{N-n}{n_d N} s'^2_{yd} .$$

5.3.2 - Sondagem aleatória estratificada

No caso de uma sondagem estratificada, o domínio de estudo poderá atravessar os estratos definidos *a priori*.



No que se segue, supõe-se que em cada estrato é efectuada uma sondagem aleatória simples sem reposição.

Define-se:

$$N_d = \sum_{h=1}^{H} N_{hd}$$

$$\mu_d = \sum_{h=1}^{H} \frac{N_{hd}}{N_d} \mu_{hd} =$$

$$= \frac{\sum_{h=1}^{H} \frac{N_h}{N} \mu_{zh}}{\sum_{h=1}^{H} \frac{N_h}{N} \mu_{xh}} = \frac{\mu_z}{\mu_x}$$

onde

$$Z_{hi} = y_{hi} 1_{i \in dh},$$
$$Z_{hi} = 1_{i \in dh}.$$

Define-se ainda

$$\overline{Z}_s = \sum_{h=1}^{H} \frac{N_h}{N} \overline{z}_h,$$

$$\overline{X}_s = \sum_{h=1}^{H} \frac{N_h}{N} \overline{x}_h.$$

Um estimador da média no domínio é dado por

$$\overline{Y}_{ds} = \frac{\overline{Z}_s}{\overline{X}_s}.$$

O valor esperado e erro quadrático médio de $\bar{Y}_{ds}$ derivam-se facilmente a partir das expressões relativas à estimação de um quociente numa amostragem estratificada.

Tem-se então

$$E(\bar{Y}_{ds}) = \mu_d + \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h n_h} \right) \left(\frac{\frac{N_{hd}}{N_h}(1 - \frac{N_{hd}}{N_h})(\mu_d - \mu_{hd})}{\left(\frac{N_d}{N} \right)^2} \right)$$

e

$$EQM(\bar{Y}_{ds}) = \left(\frac{N}{N_d} \right)^2 \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h n_h} \right) \left(\frac{N_{hd}}{N_h} \sigma'^2_{hd} + \frac{N_{hd}}{N_h}(1 - \frac{N_{hd}}{N_h})(\mu_{hd} - \mu_d)^2 \right)$$

Verifica-se que embora cada $\bar{Y}_{hd}$ seja um estimador aproximadamente centrado de μ_{hd}, $\bar{Y}_{ds}$ é estimador enviesado da média no domínio μ_d.

No caso da amostragem estratificada proporcional, o estimador da média no domínio vem

$$\bar{Y}_{ds} = \sum_{h=1}^{H} \frac{n_{hd}}{n_d} \bar{Y}_{hd} = \frac{1}{n_d} \sum_{i \in s_d} y_i .$$

O valor esperado e o erro quadrático médio do estimador aparecem numa forma um mais simplificada:

$$E(\bar{Y}_{ds}) = \mu_d + \frac{N-n}{Nn}\sum_{h=1}^{H}\frac{N_h}{N}\left(\frac{\frac{N_{hd}}{N_h}(1-\frac{N_{hd}}{N_h})(\mu_d-\mu_{hd})}{\left(\frac{N_d}{N}\right)^2}\right),$$

$$EQM(\bar{Y}_{ds}) = \frac{N-n}{Nn}\left(\sum_{h=1}^{H}\frac{N_{hd}}{N_d}\left(\frac{S_{hd}^2}{\frac{N_d}{N}} + \frac{(1-\frac{N_{hd}}{N_h})}{\frac{N_d}{N}}(\mu_{hd}-\mu_d)^2\right)\right).$$

Em particular, verifica-se que em qualquer dos casos o enviesamento será nulo se $\mu_{hd} = \mu_d, \forall h$. É assim desejável, do ponto de vista do enviesamento, que os μ_{hd} sejam o menos contrastados possível, ou seja, que a média da variável de interesse no domínio seja semelhante em todos os estratos.

Embora $\bar{Y}_{ds}$ seja um estimador enviesado, verifica-se facilmente que conduz a uma maior precisão do que aquela que se poderia obter numa amostragem simples com a mesma dimensão.

De facto, comparando os erros quadráticos médios dos estimadores da média num domínio resultantes uma estratificação proporcional, e de uma amostra não estratificada, resulta que:

$$EQM(\bar{Y}_d) - EQM(\bar{Y}_{ds}) = \frac{N-n}{Nn} \sum_{h=1}^{H} \frac{N_{hd}}{N_d} \frac{\frac{N_{hd}}{N_d}}{\frac{N_d}{N}} (\mu_{hd} - \mu_d)^2 > 0 .$$

Trata-se de uma quantidade que é sempre positiva, pelo que está à partida assegurada uma maior precisão para o estimador resultante da estratificação proporcional.

O ganho da estratificação relativamente à sondagem simples, em termos de precisão, é tanto maior quanto os μ_{hd} sejam contrastados. Tal não será de estranhar, pois afinal é este o princípio da estratificação.

Reexaminando o enviesamento tendo em conta a exigência de contraste dos μ_d, compreende-se que será preferível ter $\frac{\frac{N_{hd}}{N_h}(1-\frac{N_{hd}}{N_h})}{\left(\frac{N_d}{N}\right)^2}$ pequenos, pelo que os $\frac{N_{hd}}{N_h}$ e $\frac{N_d}{N}$ deverão estar, tanto quanto possível, perto de um.

Consequentemente, o risco de enviesamento coloca problemas nos estudos que dizem respeito a subp-opulações pouco numerosas. Se em cada estrato o domínio cobrir uma parte significativa dessa sub-população, o enviesamento será reduzido, já que o problema da estimação da média no domínio se aproxima da estimação de uma média no âmbito de uma amostragem estratificada.

Exemplo

Considere-se um concelho com 10000 habitantes onde se aproxima uma eleição municipal. O partido PX encomenda uma sondagem sobre uma amostra de 1000 munícipes, para saber a percentagem de votos que lhe são favoráveis.

Para além do total de intenções de voto, o partido está interessado em conhecer a sua penetração junto de determinados grupos etários.

O concelho é constituído por freguesias de natureza urbana e outras de natureza rural, que formam duas sub-populações com 6400 e 3600 munícipes, respectivamente.

Sondagem aleatória simples

É retirada uma amostra aleatória com probabilidades iguais e sem reposição.

O número de respondentes e a proporção de intenções de voto observados em cada uma das classes etárias são os seguintes:

Classe etária	n	$\hat{p}_d$
18-29	67	0.642
30-39	102	0.598
50-64	242	0.380
65 ou +	189	0.339

A proporção observada na amostra é usada com estimador da proporção nas populações.

A sua variância estimada é dada por:

$$\hat{V}(\hat{p}_d) = \frac{N-n}{n_d N} s'^2_{yd} \approx \frac{N-n}{n_d N} \hat{p}_d \hat{q}_d .$$

Os resultados obtidos para cada uma das classes etárias são os seguintes:

Classe etária	$\hat{V}(\hat{p}_d)$
18-29	0.0033
30-39	0.0022
50-64	0.00092
65 ou +	0.0011

Sondagem aleatória estratificada proporcional

É retirada uma amostra independente em cada uma das duas regiões com taxa de sondagem constante. A dimensão dos estratos e das amostras é:

Estrato	N_h	n_h
Urbano	6400	384
Rural	3600	216

As proporções de intenções de voto observadas em cada uma das classes etárias são as seguintes:

Classe etária	$\hat{p}_d$
18-29	0.575
30-39	0.605
50-64	0.427
65 ou +	0.290

Estas são tomadas com estimativas das proporções nas sub-populações.

Recorde-se que os erros quadráticos médio dos estimadores são aproximadamente dados por

$$E\hat{Q}M(\hat{p}_d) \approx \frac{N-n}{Nn}\left(\sum_{h=1}^{H}\frac{N_{hd}}{N_d}\left(\frac{\hat{p}_{hd}\hat{q}_{hd}}{\frac{N_d}{N}}+\frac{(1-\frac{N_{hd}}{N_h})}{\frac{N_d}{N}}(\hat{p}_{hd}-\hat{p}_d)^2\right)\right)$$

Para proceder à sua estimação recorre-se à seguinte informação relativa às dimensões das sub-populações.

Classe etária	Urbano	Rural
18-29	800	200
30-39	1600	400
50-64	2500	1500
65 ou +	1500	1500

Observam-se ainda as proporções de intenções de voto para os mesmos grupos, dadas no quadro seguinte:

Classe etária	Urbano	Rural
18-29	0,755	0,100
30-39	0,734	0,000
50-64	0,644	0,082
65 ou +	0,510	0,044

Os EQM estimados para os 4 domínios considerados vêm aproximadamente:

Classe etária	$E\hat{Q}M(\hat{p}_d)$
18-29	0.00005
30-39	0.00007
50-64	0.00014
65 ou +	0.00010

CAPÍTULO VI - MODELOS DE ESTIMAÇÃO EM CONTEXTO DE NÃO RESPOSTAS

Resumo

Neste texto apresenta-se algumas estratégias para se lidar com as não-respostas. Como veremos, o enviesamento cresce com a taxa de não-respostas. É difícil contudo, encontrar medidas objectivas do enviesamento, embora seja relativamente simples quantificar a extensão das não-respostas. Para minimizar os seus efeitos, apresentam-se estimadores que se apoiam em modelos de resposta.

1. Introdução

(i) O problema das não-respostas existe em quase todos os inquéritos[2], contudo a sua extensão e os seus efeitos variam de inquérito para inquérito. Em regra, a qualidade das estimativas produzidas dependem bastante do tratamento que os seus responsáveis são capazes de implementar para minimizar os efeitos das não respostas. Tornou-se assim habitual em organismos credenciados de estatística, usar como indicador da boa realização do inquérito a taxa de respostas, o que só realça a importância do assunto. Contudo,

[2]A amostragem como processo para obter informação relativa a um universo, não é um processo exclusivo das ciências sociais e humanas onde em regra o uso do inquérito se tornou um instrumento privilegeado de observação. A amostragem também se realiza no contexto dos processos físicos e industriais, como processo alternativo para a colheita de informação de um Universo. Mas é na verdade a possibilidade não-resposta, ou não-observação, que constitui um dos elementos distintivos da amostragem nestas ciências, uma vez que em muitos dos processos referidos (por exemplo: controle industrial da qualidade) raramente se depara com a impossibilidade de se observar a amostra gerada.

como já referimos mesmo nos "melhores" inquéritos e não obstante todos os esforços, não é possível evitar as não-respostas, tornando-se por esse motivo numa questão incontornável dos inquéritos por amostragem. O tratamento das não-respostas visa controlar a **extensão** do fenómeno, assim como os seus **efeitos** na estimação.

(ii) Acresce a este facto que em inquéritos que se repetem no tempo (por exemplo: painel) a taxa de respostas poder variar consideravelmente ao longo do tempo, o que acrescenta à lista de problemas enunciados alguma dificuldade em comparar os resultados obtidos nas sucessivas ocasiões em que o inquérito é realizado, se não houver um tratamento adequado.

(iii) Para contornar os efeitos perniciosos das não respostas, várias técnicas têm sido desenvolvidas, que se podem classificar numa das seguintes famílias:

1- Subamostragem dos Não-Respondentes
2- Aleatorização de Respostas
3- Modelização do Mecanismo de Respostas
4- Imputação de Respostas

(iv) Ao longo deste texto, pretende-se **descrever o problema das não respostas**, fornecer medidas para **avaliar a extensão do fenómeno** e fornecer uma panorâmica das técnicas que podem ser utilizadas para se proceder à **estimação em contexto das não respostas**. Como subproduto espera-se que o leitor fique na posse de um quadro, que permita avaliar conscienciosamente as situações de não-respostas e os respectivos modelos de tratamento. A modelização do mecanismo de resposta é o procedimento técnico que

permite, em tempo de estimação, incorporar e minimizar as contrariedades do processo de resposta mediante o uso das potencialidades da amostragem em duas fases. Na verdade as não respostas podem ser encaradas como o remanescente de uma subamostragem. Dificilmente o problema será completamente resolvido, mas apenas consideravelmente minimizado, o que aliás é soberbamente sintetizados pela seguinte afirmação: *"...sampling practitioners do not believe that the nonresponse models on which their adjustements are based hold exactly: they simply hope that they are improvements on the model of data missing at random over the total population implicitly assumed in estimating means with the 'do nothing' procedure." Kalton, G. in "Model in the Practice of Survey Sampling"-Int.Statist.Rev.51 (1983)*

2. Caracterização do Conceito de Não-Resposta

(i) Seleccionada uma amostra *s*, diz-se que nos encontramos perante um **problema de não-respostas**, desde que exista pelo menos um dos indivíduos da amostra[3], que não forneceu toda a informação (relativa às variáveis a observar por inquérito) que dele se pretendia.

(ii) O objectivo de um inquérito é observar um conjunto de variáveis consideradas de interesse, através de questões adequadamente formuladas para o efeito. Genericamente,

[3]Assume-se, obviamente que a base de amostragem caracteriza correctamente a população, isto é, não há erros de cobertura. Com efeito, a existência de fenómenos de natalidade e mortalidade em algumas populações, traduzem-se em erros de cobertura da base de amostragem, que levam por vezes a que na fase de recolha de informação, seja caracterizado como uma não-resposta, quando na verdade se trata de um problema de **sobre-cobertura** da base de amostragem.

pretende-se observar q variáveis, através de um inquérito, isto é:

$$\mathbf{y} = \left(y_1, \ldots, y_q\right) \tag{2.1}$$

Os responsáveis do inquérito, solicitam a resposta às questões formuladas aos indivíduos que constituem a amostra. Se o indivíduo $k \in s$ responder a todas as perguntas, teremos um **vector de respostas completo** para esse indivíduo, isto é:

$$\mathbf{y}_k = \left(y_{1k}, \ldots, y_{qk}\right) \quad k \in s \tag{2.2}$$

(iii) Dizemos que o inquérito foi **completamente respondido** (*full response*) , se todos os inquiridos responderem a todas as questões colocadas, isto é:

$$\mathbf{y}_k = \left(y_{1k}, \ldots, y_{jk}, \ldots, y_{qk}\right) \quad \forall k \in s$$

y_{jk} – resposta do individuo k, à j - ésima pergunta do inquérito (2.3) .

Neste caso, teremos uma **matriz de respostas de dimensão $n_s x q$ completamente preenchida**. Em todos os outros casos, estamos perante a existência de não-respostas, isto é, de casos em que a matriz de respostas de dimensão $n_s x q$ está **incompletamente preenchida**, ou dito de outro modo, com **valores em falta**[4].

---

[4]A designação de **missing values**, encontra-se bastante vulgarizada na língua inglesa. Informaticamente, os valores em falta são representados por "nulos" ou "brancos", com o significado óbvio, de valor a não

(iv) As razões que dão origem às não-respostas, são muito diversas e afectam diferenciadamente os inquéritos consoante a entrevista seja conduzida directamente, por via telefónica ou por via postal:

1- Na inquirição directa ou telefónica[5], a recusa em responder a todas ou a algumas questões, a dificuldade em encontrar os inquiridos na residência, ou mesmo o desconhecimento da língua (ou, linguagem) do inquirido podem significar, à posteriori, a ausência de resposta por parte de alguns elementos da amostra.

2- Na inquirição por via postal, como é obvio, todas estas dificuldades se agravam com a fraca propensão (redução do estímulo à resposta) para responder da parte dos elementos da amostra. Em geral, em inquéritos semelhantes, a realização por via postal apresenta taxas de respostas inferiores.

3- À posteriori o registo de dados revela-nos, por vezes, que por erro de digitação, por se tratar de um valor inaceitável ou ainda por ilegibilidade do valor, a

---

processar (ou a processar com regras próprias), o que remove as dificuldades de representação matricial das respostas. De outro modo, encontrar-se-ia vedada a representação matricial, em casos de não-respostas, como facilmente se percebe. Mais exactamente, as regras de processamento dos valores em falta podem (e devem) ser fixados de acordo com princípios ou critérios de natureza estatística, o que se encontra facilitado nesta representação.

[5]A audiometria, recorre de dispositivos próprios, para o estudo de audiência de meios, contudo o suporte técnico de observação, são as "linhas de comunicação telefónica".

resposta à questão não pode ser considerada e mais uma vez (objectivamente) nos encontramos perante uma não-resposta.

(v) Designemos o **conjunto de indivíduos que responderam a uma dada questão** j, por

$$r_j = \left\{k : k \in s \wedge y_{jk} \text{ conhecido}\right\} \qquad (2.4)$$

o que nos permite definir no máximo q conjuntos (distintos) de respondentes (um para cada questão). Todos eles são obviamente subconjuntos da amostra s, isto é, o resultado de uma "subamostragem" em s.

(vi) A existência de não-respostas, coloca dificuldades à estimação dos parâmetros de interesse. Estas dificuldades derivam (sobretudo) de três razões essenciais:

1 - Em regra são **desconhecidos os mecanismos que dão origem às não-respostas**, ou melhor, como são gerados os conjuntos de respondentes, isto é:

$p_j\left(r_j \mid s\right)$ – probabilidade de r_j ser o conjunto dos respondentes à j - ésima questão do inquérito, para uma dada amostra s.

$p_j\left(\cdot \mid s\right)$ – lei de probabilidade desconhecida

Para contornar, as dificuldades apresentadas, são geralmente consideradas algumas hipóteses razoáveis de tratamento, a saber:

1ª Hipótese: Admite-se que o comportamento de cada indivíduo da amostra relativamente ao inquérito não **depende** dos outros indivíduos da amostra, isto é:

$$p_j\left(r_j|s\right) = p_{j1}\cdots p_{jk}\cdots p_{jK}$$

p_{jk} - probabilidade de o individuo $k \in s$ responder à j - ésima questão do inquérito (2.5)

2ª Hipótese: Admite-se que há indivíduos com **características semelhantes**, isto é, com características tais que dão origem a **comportamentos semelhantes perante um inquérito**. Esta hipótese é útil para a estimação da probabilidade de responder de cada indivíduo. Com maior generalidade, diga-se que se existe informação caracterizadora dos indivíduos então, por recurso a modelos, é possível estimar para cada um a propensão de responder ao inquérito.

2 - As inferências, isto é, a qualidade da estimação depende em grande medida, da **validade das hipóteses assumidas**. A formulação de hipóteses, é não só compreensível em face do problema, como também inevitável em inquéritos por amostragem, embora seguramente indesejável. A experiência dos responsáveis e a sua sensibilidade é fundamental, para a formulação de modelos credíveis, que a seu tempo serão testados e apoiarão a estimação.

3 - A estimação basear-se-á obviamente no conjunto dos respondentes, ou dizendo de outra maneira, a construção dos estimadores deve ter em conta eventuais cenários de não respostas (resultado de uma "subamostragem"). Neste contexto, a estimação é de uma complexidade acrescida,

porque temos de nos confrontar com **um processo de amostragem, dentro do qual ocorre uma "subamostragem"**, o que a torna mais complexa, ainda que o mecanismo da "subamostragem" seja conhecido.

4 - O facto de haver usualmente **várias perguntas**, no mesmo inquérito, acresce à complexidade referida uma maior dimensão, uma vez que, o comportamento dos inquiridos (isto é, as não-respostas) podem diferenciar-se de questão para questão.

3. Extensão das Não Respostas.

(i) Em consequência das não respostas, o enviesamento aumenta (como adiante mostraremos) e uma vez que as hipóteses formuladas para o seu tratamento podem estar de algum modo distorcidas a quantificação exacta dos seus efeitos, é em regra difícil. Sendo difícil a quantificação dos efeitos, não o é a avaliação da sua extensão, isto é, saber "quantos inquiridos" não responderam ou "quanta informação" se perdeu não é complicado. Por isso, em geral, a divulgação de estimativas é, muitas vezes, acompanhada de uma medida de não respostas para que os utilizadores possam ajuizar (subjectivamente) da credibilidade dos resultados divulgados. A importância de um indicador deste tipo reside ainda no facto de que tanto o enviesamento como a imprecisão do processo de estimação, em regra aumentarem, com a diminuição da taxa de respostas.

(ii) Neste contexto, convêm precisar, alguns conceitos relativos às não-respostas:

1 - O elemento k da amostra constitui um **não-respondente ao inquérito** (*unit nonresponse*), se se recusou responder a todas as questões do inquérito, isto é:

$$k \text{ é um não - respondente ao inquérito } \Leftrightarrow k \notin r_u \qquad (3.1)$$
$$r_u = r_1 \cup \ldots \cup r_j \cup \ldots \cup r_q \text{ (unit response set)}$$

2 - O elemento k da amostra constitui um **não respondente a questões** (*item nonresponse*), se se recusou a responder a alguma (a pelo menos uma) das questões colocadas, isto é:

$$k \text{ é um não - respondente a questões } \Leftrightarrow k \notin r_c \qquad (3.2)$$
$$r_c = r_1 \cap \ldots \cap r_j \cap \ldots \cap r_q \text{ (item response set)}$$

3 - O elemento k da amostra é um **não respondente à questão *j*** (*item j nonresponse*), se se recusou a responder a essa questão, isto é:

$$k \text{ é um não - respondente à questão j } \Leftrightarrow k \notin r_j \qquad (3.3)$$
$$r_j \quad \text{(item } j \text{ response set)}$$

4 - Ficam assim definidos os seguintes conjuntos notáveis:

$s - r_u$ - Conjunto de não - respondentes ao inquérito(unit nonresponse set)

$r_u - r_c$ - Conjunto de não - respondentes a questões (item nonresponse set)
{ dos que responderam ao inquérito }

$r_u - r_j$ - Conjunto de não - respondentes à questão j (item j nonresponse set)
{ dos que responderam ao inquérito }

Note-se que o conjunto de referência, é o conjunto r_u , embora nada obste a que se convencione um outro conjunto para esse efeito.

(iii) Um caso de grande valor pedagógico[6], é o caso em que os inquiridos se comportam do mesmo modo em relação a todas as questões, isto é:

$$r = r_1 = ... = r_j = ... = r_q \Rightarrow r_u - r_j = \varnothing \qquad (3.4)$$

(iv) Em face do exposto, duas medidas da capacidade de obter a informação desejada são utilizadas habitualmente, a saber:

1- A probabilidade de um indivíduo escolhido aleatoriamente[7], de entre aqueles que integram a amostra ter respondido[8] ao inquérito (realizado em condições conhecidas). Trata-se obviamente de uma probabilidade condicional, isto é, dado que o indivíduo foi seleccionado para a amostra. A sua estimação à posteriori, resulta de:

---

[6] Que será bastante usado nesta exposição, em parte, para simplificar a exposição.

[7] Sendo a selecção aleatória efectuada de modo a garantir para todos os indivíduos a equiprobabilidade, ou seja, para efeitos de controle da qualidade da inquirição (respostas), todos os indivíduos da amostra tem a mesma importância. Trata-se de estimar a probabilidade de um indivíduo escolhido "ao acaso" na amostra, vir a responder ao inquérito.

[8] A distinção entre respondentes a uma questão determinada, a todo o questionário, ou a pelo menos parte do questionário, só será considerada quando a distinção for relevante para o desenvolvimento teórico. A adaptação das formulas para casos diversos, faz-se sem qualquer dificuldade.

$$P(k \in r|s) = \sum_{k \in s} p_{ks} \underbrace{P(k \text{ ser selececcionado} | k \in s)}_{\frac{1}{n_s}} = \frac{1}{n_s} \sum_{k \in s} p_{ks}$$

(Teor. das Prob. Totais) (3.5)

$$p_{r|s} = \frac{1}{n_s} \sum_{k \in s} y_k = \frac{n_r}{n_s} \qquad y_k = \begin{cases} 0 \text{ se não respondeu} \\ 1 \text{ se respondeu} \end{cases}$$

$$E\{p_{r|s}|s\} = \frac{1}{n_s} \sum_{k \in s} E\{y_k|s\} = \frac{1}{n_s} \sum_{k \in s} p_{ks} = P(k \in r|s)$$

Como se percebe as características do desenho de amostragem, no espírito desta medida, não influenciam o comportamento dos inquiridos, daí que as taxas de resposta seriam supostamente homogéneas. Em concreto, admite que a perda de informação é indiferenciada, isto é, não se relaciona com o desenho de amostragem razão que leva à consideração da amostra como grupo de referência homogéneo.

2- A probabilidade de num desenho de amostragem, um indivíduo seleccionado aleatoriamente[9] na população, pertencer ao grupo dos que não se recusam (ou não recusariam) a responder ao inquérito. De modo análogo, esta probabilidade pode ser "à posteriori" estimada, com base na teoria de amostragem, e em particular através do clássico estimador π de Horvitz-Thompson:

---

[9] A selecção aleatória é efectuada admitindo probabilidades iguais, isto é, para efeitos de controle da qualidade da inquirição (desenho + respostas), todos os indivíduos da população tem a mesma importância. Trata-se de estimar a probabilidade de um indivíduo escolhido "ao acaso" na População não se recusar a responder ao inquérito.

$$\hat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} \qquad E\left[\hat{t}_\pi\right] = t = \sum_k y_k \tag{3.6}$$

π_k – probabilidade de $k \in s$ pertencer à amostra

assim podemos derivar:

$$P(k \in r) = \sum_{k \in U} p_{ks} \underbrace{P(k \text{ ser seleccionado})}_{\frac{1}{N}} = \frac{1}{N} \sum_{k \in U} p_{ks}$$

(Teor. das Prob. Totais)

(3.7)

$$p_r = \frac{1}{N} \sum_{k \in U} \frac{y_k I_k}{\pi_k} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k} = \frac{1}{N} \sum_{k \in r} \frac{1}{\pi_k}$$

$$y_k = \begin{cases} 0 \text{ se não respondeu} \\ 1 \text{ se respondeu} \end{cases}$$

$$E\{p_r\} = \frac{1}{N} \sum_{k \in U} \frac{1}{\pi_k} E\left\{ \overbrace{\underbrace{E\{y_k | k \in s\}}_{p_{ks}} I_k}^{\text{independentes}} \right\}$$

$$= \frac{1}{N} \sum_{k \in U} \frac{1}{\pi_k} p_{ks} \underbrace{E\{I_k\}}_{\pi_k} = P(k \in r)$$

Se atendermos ao facto de que em desenhos de dimensão variável, existe um estimador alternativo (baseado no ratio) mais eficiente que o estimador π, então podemos modificar o estimador para aumentar a sua eficiência, isto é:

$$p_r = \frac{\sum_{k \in s} \frac{y_k}{\pi_k}}{\sum_{k \in s} \frac{1}{\pi_k}} = \frac{\sum_{k \in r} \frac{1}{\pi_k}}{\sum_{k \in s} \frac{1}{\pi_k}} \qquad (3.8)$$

De notar que, a taxa de resposta assim calculada não depende do desenho de amostragem. Note-se apenas, o facto de esta taxa assentar o seu calculo, na presunção de que o comportamento (probabilidade de responder) do inquirido não se altera de amostra para amostra (requisito de independência de p_{ks} *e* I_k), ou seja, que as condições de inquirição permanecem inalteradas de amostra para amostra.

3- Se houver uma medida da quantidade de informação (variável x) proporcionada por cada indivíduo, podemos de modo análogo estimar a percentagem de informação recolhida, isto é:

$$p_{xr|s} = \frac{\sum_{k \in r} x_k}{\sum_{k \in s} x_k} \qquad p_{xr} = \frac{\sum_{k \in r} \frac{x_k}{\pi_k}}{\sum_{k \in s} \frac{x_k}{\pi_k}} \qquad (3.9)$$

onde, a primeira fórmula representa a percentagem de informação recolhida, de entre aquela que se pretendia recolher, isto é, no contexto da amostra. A segunda formula representa a percentagem da informação caracterizadora[10] do universo recolhida, de entre a informação necessária para a caracterização adequada/desejada do Universo, isto é, o ponto de referência é o Universo.

A distinção entre medidas condicionadas e não condicionadas das respostas, esbate-se no caso de se tratarem de desenhos de amostragem com iguais probabilidades de inclusão.

(v) As medidas de não-respostas, são adequadas para caracterizar a capacidade de recolha de informação, mas dizem pouco sobre as suas causas e sobretudo acerca dos seus efeitos na estimação, que como já se referiu é um complicado problema metodológico.

4. Não-Respostas: Avaliação Empírica das Consequências

(i) Uma boa estratégia de amostragem para um dado inquérito, deve prever a existência de não respostas e conceber estratégias destinadas ao seu tratamento. Em particular, as acções a empreender podem ser de três tipos, consoante a fase em que se processa a intervenção, a saber:

1-Na fase de **planeamento**, são definidas as acções a empreender de forma a minimizar as não respostas, isto é,

[10] O valor da informação, depende também do desenho de amostragem utilizado, o que justifica a ponderação da medida de informação. O desenho de amostragem, não só determina o modo de selecção da amostra, como também **condiciona a escolha dos estimadores** adequados à caracterização do universo.

pretendendo reduzi-las a níveis insignificantes e/ou a minimizar o enviesamento. Além de algumas recomendações de "bom senso"[11], como a insistência junto do inquirido e o cuidado com a extensão do questionário, a **aleatorização das respostas** [Warner(1965)] é um dos processos que poderá ser utilizado para esse fim, sobretudo nos casos em que os inquiridos revelam relutância em responder às questões colocadas. Na verdade, um mecanismo aleatório cripta a resposta do inquirido de forma a impedir a sua descodificação, mas com a delicadeza necessária para não impedir a estimação.

2-Na fase de **recolha**, perante a constatação de que existem não-respostas, poder-se-á proceder a uma **subamostragem de não respondentes** [Hansen e Hurwitz (1946)] , uma vez que uma parcela da estimativa desejada continua desconhecida. Trata-se de um procedimento para controlar/dominar o processo de não-respostas que possibilita a estimação não-enviesada, recorrendo à estimação por subamostragem da parcela desconhecida por efeito das não-respostas. Apesar da elegância da técnica usada, dever-se-á alertar para o seu elevado custo de implementação, sobretudo quando se trata de inquéritos com alguma dimensão, uma vez que se exige o conhecimento da matriz de resposta completa da subamostra.

3-Na fase de **estimação**, perante a necessidade de produzir resultados, o conhecimento do **mecanismo** (aleatório) **das não-respostas** permite-nos contornar o enviesamento

[11] Apesar disso, nem sempre óbvias, conforme se constata da prática de algumas instituições com responsabilidade nesta área.

proveniente das não-respostas, mediante a consideração das probabilidades com que cada inquirido apresenta responde ao inquérito. A **modelização do mecanismo desconhecido de resposta** permite em função de informação auxiliar, proceder à estimação das referidas probabilidades e reduzir o enviesamento do estimador. Em certo sentido a **imputação**, é um processo semelhante, na medida em que, através do uso de informação auxiliar relevante (isto é, relacionada com a variável desconhecida e baseando nas semelhanças entre indivíduos), poder-se-á contribuir para a produção de "bons" valores a imputar (estimativas para indivíduos). É a apresentação destes mecanismos que visa a presente exposição.

(ii) As atitudes de **"bom senso"** que referimos, dirigem-se aos factores que são responsáveis pelas não-respostas, dos quais se devem referir:

1 - A selecção, formação e supervisionamento dos entrevistadores são actividades relevantes, que devem ser alvos do maior cuidado, uma vez que, em geral, se encontra na mão dos entrevistadores, a possibilidade de se conseguir a colaboração do entrevistado.

2 - As questões, isto é, a forma como elas são colocadas e a própria extensão do questionário, devem ser avaliadas de forma que o nível de não respostas, se mantenha dentro de limites aceitáveis. Paralelamente[12] a escolha do método de recolha de informação , isto é, por via postal, por telefone ou por entrevista directa, deve ser ponderada no contexto

---

[12] Mas tendo em consideração as características do questionário, nomeadamente a extensão.

das disponibilidades orçamentais, uma vez que dela derivam diferentes taxas de respostas.

3 - A relação de confiança estabelecida entre a instituição responsável pela operação e os inquiridos é fundamental, uma vez que dela resultará nomeadamente a convicção de que não será realizado qualquer uso abusivo da informação fornecida. Se assim suceder, reduzir-se-ão as razões que muitas vezes levam os inquiridos a recusar a sua colaboração.

Não há dúvidas, do ponto vista lógico, de que a forma como as instituições encaram as não-respostas, é muito importante para a qualidade da informação produzida. No nosso país, esta preocupação tem estado cada vez mais presente em algumas instituições com responsabilidades na produção estatística. Mas importa referir, que nem sempre a força da razão tem sido suficiente para a mudança de atitude de muitas instituições. Dalenius (1976), a este propósito, apresenta o caso de duas instituições responsáveis pela condução de um mesmo inquérito, na mesma região, em que uma delas teve apenas 10% de não-respostas, enquanto outra teve uma taxa de não-respostas de 50%. A diferença, resultou - segundo Dalenius - da diferença de atitude, que as duas instituições tinham perante as não-respostas.

(iii) A argumentação apresentada sugere que os efeitos das não-respostas não se limitam a uma redução da dimensão da amostra, isto é, à redução da eficiência. Pelo contrário, modificam outras propriedades dos estimadores e em especial o enviesamento. A concepção de uma estratégia de amostragem credível exige a modificação do estimador, para contornar as dificuldades apresentadas e sobretudo para o

tornar mais robusto. Sugerimos que as estimativas produzidas por esses estimadores - na ignorância do mecanismo das respostas - afastam-se, em média, significativamente do parâmetro a estimar. Adiante, justificar-se-á esta ideia empírica com argumentos formais. Por agora, limitar-nos-emos a ilustrar com dois casos para reforçar a ideia de que as não-respostas devem ser encaradas com cuidado. Uma ideia básica, para reduzir as não-respostas, é a de tentar mais do que uma insistência no contacto[13] com os inquiridos em falta, o que - em regra - contribui significativamente para a redução das não-respostas. Ao mesmo tempo, o acompanhamento das insistências permite-nos aperceber melhor os efeitos das não-respostas. Com efeito, vejamos:

1- P. Rao (1983) cita os dados de Hilgard e Payne (1944), onde se pretendia conhecer a percentagem de agregados familiares com filhos de idade inferior a 2 anos. No primeiro contacto responderam 63.5% dos inquiridos (que se encontravam na residência), dos quais 17.2% tinham crianças com idade inferior a 2 anos. Das pessoas que responderam ao segundo contacto (22.2%), só 9.5% tinham crianças nas idades referidas. Os restantes agregados responderam ao terceiro contacto, mas destes apenas 6.2% tinham crianças nas idades referidas. A explicação óbvia, parece ser a redução de mobilidade a que se encontram sujeitos os agregados com crianças com idades inferiores a 2 anos, o que facilita o contacto pelos entrevistadores no local de residência.

---

[13] Em Portugal vulgarizou-se, em alguns organismos com responsabilidade na produção estatística, a designação de insistência.

	Nº de Insistências 1	2	>2	Não Respostas	Total
Percentagem de Respostas	63.5%	22.2%	14.3%	0%	100%
Percentagem de Famílias com crianças com menos de 2 anos	17,2%	9.5%	6.2%	-	13.9%

2- Um outro exemplo, é dado por Danermark e Swenson(1987), relativo a um inquérito realizado a estudantes com idades superiores a 16 anos em estabelecimentos de ensino, com o objectivo de saber a percentagem de estudantes que já haviam experimentado drogas leves (marijuana e haxixe). No primeiro contacto foram inquiridos 89.6% dos indivíduos da amostra dos quais apenas 4.9% tinham experimentado drogas leves. No segundo contacto foram inquiridos mais 8.9%, dos quais 14.9% já haviam experimentado drogas leves. A relação (e não causalidade) existente entre o consumo de drogas leves e algum absentismo escolar, será eventualmente a razão da discrepância entre os resultados provenientes dos dois contactos.

	Nº de Insistências 1	>1	Não Respostas	Total
Percentagem de Respostas	89.6%	8.9%	1.5%	100%
Perc. de Estudantes que já experimentaram Haxixe e/ou Marijuana	4.9%	14.9%	?	?

Como dissemos, as não-respostas podem subavaliar ou sobreavaliar os parâmetros, conforme foi evidenciado nos

exemplos precedentes. Podia-se, sem grande dificuldade, acrescentar outros exemplos, evidenciadores das consequências das não respostas. As suas consequências não devem ser menosprezadas, pois delas resultam enviesamentos nos parâmetros e justificam pelo menos um esforço de insistência nos contactos. Em inquéritos, por via postal, a realização de duas a três insistências é mesmo uma norma aceite em inquéritos que prezam alguns requisitos de qualidade.

5. Utilização do Mecanismo de Respostas para a Estimação

Noção de Amostragem Quase-Probabilística

(i) A modelização do mecanismo das respostas, assenta na ideia de que se fosse conhecida a propensão (probabilidade) com que cada elemento da amostra responde ao inquérito, então poder-se-ia construir um estimador para os parâmetros da população, tirando partido da amostragem em duas fases.

(ii) Como atrás referimos, admitem-se duas hipóteses razoáveis, de forma a permitir a modelização do mecanismo das respostas, a saber:

1- A independência entre os indivíduos na decisão de responder ou não ao inquérito.

2- Que indivíduos com características semelhantes apresentam comportamentos semelhantes em relação ao inquérito.

(ii) Num primeiro momento, considere-se apenas a primeira hipótese e vamos supor conhecidas as probabilidades com que os indivíduos se dispõem a fornecer a informação

solicitada. Neste caso, de acordo com a teoria da amostragem em duas fases[14], teremos como estimador condicional do total π^* a substituir o estimador π de Horwitz-Thompson:

$$\hat{t}_{\pi^*} = \sum_{k \in r} \frac{y_k}{\pi_k \theta_k} \quad \text{a substituir} \quad \hat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} \quad \text{com} \quad E\left[\hat{t}_\pi\right] = t \quad (5.1)$$

$P(k \in s) = \pi_k$ - probabilidade do indivíduo k da população pertencer à amostra.

$P(k \in r|s) = \theta_k$ - probabilidade do indivíduo k da amostra s responder ao inquérito.

donde:

$$E_{RD}\left[\hat{t}_{\pi^*}|s\right] = \hat{t}_\pi \Rightarrow E\left[\hat{t}_{\pi^*}\right] = t \qquad (5.2)$$

(iii) O facto de a resposta (inclusão dos indivíduos na segunda fase) ser independente de indivíduo para indivíduo, aponta para uma amostragem de Poisson na segunda fase. Neste caso, como sabemos, existe um estimador mais eficiente não-informativo do que o estimador π, conhecido como o estimador alternativo, inspirado no estimador do ratio e aproximadamente não-enviesado, cuja expressão é:

[14] O uso desta notação, neste contexto pretende reforçar a ideia de que as probabilidades de inclusão da 2ª fase, são um parâmetro desconhecido θ a estimar. Não se considerará, por razões de oportunidade, as probabilidades de inclusão de 2ª ordem, assim como a estimação da variância a elas associada. A sigla RD, é usada como abreviatura de Response Distribution.

$$\hat{t}_1 = N \frac{\sum_{k \in r} \frac{y_k}{\pi_k \theta_k}}{\sum_{k \in r} \frac{1}{\pi_k \theta_k}} = N \bar{y} \qquad (5.3)$$

(iv) O caso mais simples, acontece, se todos os indivíduos da amostra apresentam a mesma probabilidade de responder. Neste caso o desenho de Poisson, reduz-se a um desenho de Bernoulli e então temos:

$$\theta_k = \theta \Rightarrow \hat{t}_1 = N \frac{\sum_{k \in r} \frac{y_k}{\pi_k \theta}}{\sum_{k \in r} \frac{1}{\pi_k \theta}} = N \frac{\sum_{k \in r} \frac{y_k}{\pi_k}}{\sum_{k \in r} \frac{1}{\pi_k}} \qquad (5.4)$$

Neste caso, admite-se implicitamente que o mecanismo de resposta não influência a estimação, como se percebe da leitura do estimador. O único efeito situar-se-á a nível da variância, que aumenta com a redução da dimensão da amostra. Formalmente, tudo está correcto, **mas convenhamos que o modelo apresenta uma fraca adesão à realidade**.

(v) No caso do modelo de Bernoulli coincidir com o verdadeiro mecanismo de respostas, não haverá inconveniente em o utilizar. O enviesamento do estimador resultará apenas da sua inspiração no ratio e que, como se sabe, em amostras com alguma dimensão é aproximadamente nulo. Os problemas põem-se naturalmente, quando o modelo se afasta da realidade. Dois modelos podem ser então, usados para análise:

1- O verdadeiro modelo é de Poisson. Neste caso, o enviesamento será:

$$B\left[\hat{t}_1\right] = E\left[\hat{t}_1\right] - t \cong N\frac{\sum_k y_k\theta_k}{\sum_k \theta_k} - t = \frac{N\overline{y\theta}_U}{\overline{\theta}_U} - \frac{N\overline{y}_U\overline{\theta}_U}{\overline{\theta}_U}$$

$$\text{Como } S_{y\theta U} = \frac{\sum_k \left(y_k - \overline{y}_U\right)\left(\theta_k - \overline{\theta}_U\right)}{N-1} \qquad R_{y\theta U} = \frac{S_{y\theta U}}{S_{yU}S_{\theta U}}$$

$$B\left[\hat{t}_1\right] = (N-1)\frac{S_{y\theta U}}{\overline{\theta}_U} = \frac{t(N-1)R_{y\theta U}cv_{yU}cv_{\theta U}}{N}$$

donde:

$$RB\left[\hat{t}_1\right] = \frac{B\left[\hat{t}_1\right]}{t} \cong R_{y\theta U}cv_{yU}cv_{\theta U}$$

isto é

$RB\left[\hat{t}_1\right]$ depende de $R_{y\theta U}$

O que mostra que o problema do enviesamento é tanto mais importante, quanto maior for a **relação entre os valores da variável e a probabilidade de responder**.

2- O mecanismo de resposta é determinístico, isto é, a população é estratificada em dois grupos, um que responde sempre (com probabilidade 1) e outro que nunca responde (com probabilidade 0). O enviesamento, será então:

U_1, U_2 - Estractos de respondentes e não - respondentes com dimensões N_1, N_2 respectivamente

$$B\left[\hat{t}_1\right] \cong N_2(\bar{y}_{U_1} - \bar{y}_{U_2})$$

O que mostra, neste caso, que o enviesamento aumenta com a **diferença das médias dos estratos** e com a **dimensão do estrato dos não-respondentes**. Este resultado, alerta-nos para a necessidade de conceber estratégias, que reduzam o peso do estrato de não-respondentes.

Se na teoria base da amostragem, a ênfase é colocada na redução da dispersão, por motivos óbvios a presença das não-respostas, **recoloca o problema do enviesamento**. A ideia é obviamente, de procurar conhecer o mecanismo de respostas e em face do conhecimento deste mecanismo eliminar ou pelo menos reduzir o enviesamento.

(vi) Durante muitos anos, o **modelo determinístico** foi a base para o tratamento das não respostas, seguindo uma sugestão de Cochran(1977). Assente neste modelo, nas décadas de 60 e 70, muitos trabalhos de tratamento de não respostas foram desenvolvidos nos Institutos de Estatística. Mas como confessava Cochran, a ideia baseava-se numa supersimplificação da realidade. Ainda é esta ideia, que remotamente inspira, o processo de subamostragem dos não-respondentes. O inconveniente deste processo - como referimos - é o seu custo e a sua difícil aplicação a amostragens de grande dimensão. São estes inconvenientes, que nos remetem para a pesquisa de outras soluções. Da sua exploração, ficam duas ideias importantes para a redução do enviesamento, a **necessidade de reduzir a dimensão do**

estrato das não respostas e que o **enviesamento é proporcional à diferença das médias dos estratos**[15].

(vii) Uma representação mais fiel da realidade, assenta na hipótese de que o processo de respostas é aleatório e não determinístico. Como mostramos, a ideia de que todos indivíduos possuem o mesmo mecanismo aleatório de resposta, conduz-nos à amostragem de Bernoulli na segunda fase. É a versão mais simples do modelo aleatório. Neste caso a estimação, como vimos, considera que das não-respostas apenas resulta uma redução da dimensão da amostra. Em termos práticos, equivale isto a dizer, que as não-respostas mantém inalteradas as propriedades do não-enviesamento, o que em geral, não é aceitável. Sociologicamente, corresponde a assumir que não existem características individuais (ou diferenças de comportamento), que aumentem ou reduzam a propensão a responder e sobretudo que essa propensão não se encontra correlacionada com a variável de interesse. Da exploração efectuada, há que reter a ideia de que o enviesamento resulta, da associação (correlação) que pode existir entre o mecanismo de respostas e os valores da variável, o que pode comprometer os valores observados, sobrevalorizando o perfil dos respondentes.

(viii) Em qualquer dos casos considerados, o verdadeiro mecanismo de respostas é na prática desconhecido. As alternativas, são "impor" um dos modelos sugeridos ou em face da informação disponível, estimar as probabilidades de respostas. A estimação parece a solução mais sensata e que

[15]Se existir uma associação (correlação) entre o valor da variável de interesse e a resposta (ou não) ao inquérito, então o modelo determinístico, permite dizer que a estimação efectuada através dos respondentes é enviesada. Esta constatação, é retomada no modelo aleatório.

mais se aproximará do verdadeiro modelo. É esta a abordagem, que adiante desenvolveremos. Importa no entanto reconhecer, que o uso desta abordagem afasta-nos dos conceitos da amostragem probabilística, uma vez que se desconhece o desenho de amostragem que dá origem à "amostra observada". Autores existem que sugerem a designação de "**amostragem quase-probabilística**"[16], para os métodos em que o desenho de subamostragem só se torna conhecido por recurso à estimação.

(ix) Ao contrário do que à primeira vista se poderia julgar, como já se percebe, **a estimação no contexto das não-respostas não desvaloriza o papel da teoria da amostragem, mas antes exige a concepção de modelos robustos e resistentes**. O seu desenvolvimento assenta obviamente, na amostragem em duas fases. Em todo o caso, o problema do enviesamento[17] nunca se encontrará completamente resolvido, mas apenas substancialmente reduzido, isto é, quando muito minimizado, face à informação disponível.

(x) Geralmente, mas não necessariamente, a estimação do mecanismo de resposta é precedido da imputação de "bons valores" para os indivíduos que responderam apenas

---

[16] Os trabalhos de Oh e Scheuren(1983) e Plateck e Gray (1983) seguem esta abordagem, não apenas para a resolução do problema das não-respostas, mas visando também a resolução de outros erros não-amostrais.

[17] É o que refere, Kalton(1983) e Kalton e Kasprzyk(1986), ao considerarem que o conhecimento do verdadeiro mecanismo de resposta depende não só da validade das hipóteses formuladas, como também dos procedimentos de estimação.

parcialmente ao inquérito, isto é, apenas a algumas questões do inquérito. Nestes casos, a imputação será obviamente uma estratégia a considerar.

6. Modelização do Mecanismo de Respostas
Uma Estratégia Realizável de Estimação: Grupos de Resposta Homogéneo

(i) A sofisticação estatística, nem sempre é compatível, com a necessidade de produzir estimativas rápidas, o que aconselha em muitos casos o uso técnicas robustas mas simultaneamente simples de estimação. Neste sentido, a estratificação dos elementos da amostra, em grupos que apresentem comportamentos homogéneos de resposta, torna-se um modelo simples, mas poderoso de estimação. O modelo RHG (response homogeneity group) é caracterizado do seguinte modo:

1. A amostra é repartida em H_s grupos distintos, com comportamentos hómogeneos de resposta:

$$s = s_1 \cup \ldots \cup s_h \cup \ldots s_{H_s} \quad s_h \cap s_{h'} = \varnothing \; \forall h \neq h'$$

2. As probabilidades de inclusão são uniformes em cada grupo, isto é:

$$\pi_{k|s} = P(k \in r|s) = \theta_{hs} \quad \forall k \in s_h$$

Supõem-se agora - como se percebe - que quem respondeu ao inquérito, o fez a todas as questões do inquérito. Esta suposição, não retira a generalidade ao modelo e é geralmente contornada como referimos através da imputação. O objectivo é sobretudo de se reduzir a complexidade da estimação.

(ii) Deve-se notar, em primeiro lugar que para efeitos de estratificação da amostra, se usa não apenas a informação prévia à amostragem, mas também a informação relativa à amostra concreta (da 1ª fase) na qual se inclui a informação relativa às condições em se processou a recolha de informação, e que é susceptível de ter influenciado o mecanismo de resposta. Em particular, diga-se que:

1. A formação de grupos, pode variar de amostra para amostra[18], de acordo com a informação disponível para a caracterização do mecanismo de resposta.

2. A estratificação, pode (e deve) levar em conta os meios e as dificuldades ocorridas no "trabalho de campo", como o número de entrevistadores e/ou o seu perfil, que se traduzem no processo de resposta, ou seja, as condições de inquirição.

O sucesso do modelo, assenta na capacidade e astúcia do estaticista, em formar grupos com mecanismos de resposta homogéneos. Assim a identificação dos factores que condicionam as respostas, desempenha um papel determinante, nas estratégias de amostragem quase-probabilistica.

(iii) No caso de o modelo ser verdadeiro, as probabilidades de inclusão, poderão ser estimadas, com base nas frequências observadas em cada grupo homogéneo. Neste caso, sob a hipótese de validade do modelo, há a considerar:

[18] Pode variar de amostra para amostra, mas não numa mesma amostra sob condições dadas de inquirição, isto é, as condições de inquirição ficam determinadas antes do processo de resposta.

1. Em cada um dos H_s grupos distintos, foram solicitadas respostas a:

(6.3)

$$|s_h| = n_h \text{ inquiridos e houve } |r_h| = m_h \text{ respondentes}$$

$$\mathbf{m} = (m_1,\ldots,m_h,\ldots,m_{H_s})$$

2. As probabilidades de inclusão são estimadas de acordo com o modelo e com as frequências observadas: (6.4)

$$\pi_{k|s,m} = P(k \in r|s,\mathbf{m}) = \frac{m_h}{n_h} = f_h$$

$$\forall k \in s_h \text{ pois } E_{ARD}(m_h) = n_h\theta_{hs} \qquad n_h \geq 1$$

As probabilidades de inclusão da 2ª fase são agora condicionadas à validade do modelo, isto é, à distribuição das respostas assumida (*ARD*). Este modelo foi detalhadamente estudado por Särndal e Swenson (1987). A sua utilização pode ser feita, com ou sem o uso de variáveis auxiliares na fase de estimação, como adiante veremos.

(iv) Um modelo, com inspiração algo semelhante, deve-se a Oh e Scheuren (1983) conhecido, como Modelo de Subpopulações com Mecanismos de Resposta Uniforme. A grande diferença em relação ao modelo preconizado, reside no facto de **se realizar a estratificação da amostra, num momento prévio à amostragem**, isto é, sem atender à amostra particular e às condições de realização do inquérito.

(v) Autores de inspiração Bayesiana[19], efectuam a distinção entre **mecanismo de resposta ignorável e não-ignorável**, pretendendo com isso notar que se o mecanismo de resposta é não-correlacionado com a variável de interesse, então o

[19] É o caso dos estimadores apresentados por Rubin (1976, 1983 e 1987) e Little e Rubin (1987).

conhecimento do mecanismo é irrelevante para a estimação. Não se pondo em causa, a ideia atrás referida, que o mecanismo de resposta só é relevante quando correlacionado com a variável de interesse, julga-se problemático assumir "à priori" a sua irrelevância, isto é, sem atender a informação que o confirme.

(vi) Estes resultados (6.4 e 5..3) , permitem construir os estimadores condicionais não enviesados[20] e determinar as suas propriedades, usando a técnica de amostragem estratificada de Bernoulli na 2ªfase, a saber:

$$\hat{t}_{c\pi^*} = \sum_{h=1}^{H_s} \sum_{k \in r} \frac{\vee y_k}{\pi_{k|s,m}} = \sum_{h=1}^{H_s} f_h^{-1} \sum_{k \in r} \vee y_k \qquad (6.5)$$

(vii) Importa notar, que tudo assenta na hipótese de que o modelo assumido é coincidente com o verdadeiro mecanismo de resposta, caso em que, o estimador[21] é não enviesado, pois

$$E\left[\hat{t}_{c\pi^*}|s\right] = E_m E_{RD}\left[\hat{t}_{c\pi^*}|s,m\right] = E_m E_{RD}\left[\sum_{h=1}^{H_s} f_h^{-1} \sum_{k \in r} \vee y_k |s,m\right] = \qquad (6.6)$$

$$= E_m\left[\sum_{h=1}^{H_s} \sum_{k \in s} \vee y_k |s\right] = \hat{t}_\pi$$

Que tem o efeito de corrigir, a dimensão esperada da amostra através da dimensão da amostra observada.

(viii) Este estimador quando aplicado a desenhos usuais - para efeitos de computação - apresenta expressões menos elegantes, mas que por serem o resultado de um trabalho

[20] Se o modelo assumido for verdadeiro.

[21] Refira-se, que neste caso a notação é pouco clara e confunde o estimador e a estimativa.

adequado, reduzem a complexidade computacional. Por exemplo, no caso do desenho SI (amostragem aleatória simples sem reposição na 1ª fase), temos:

$$\hat{t}_{c\pi^*} = N\left[\underbrace{\frac{\sum_{h=1}^{H_z} n_h \bar{y}_{r_h}}{n}}_{\hat{\bar{y}}_U}\right] = N\,\hat{\bar{y}}_U \qquad (6.7)$$

(ix) Uma recomendação efectuada, sempre que se trabalha em contexto das não respostas, é a de se usar sempre que possível estimadores informativos, como o da diferença ou o da regressão. A grande vantagem de uma estratégia assim definida é de aumentar não só a eficiência, como também a resistência a percalços de amostragem, mesmo que o modelo assumido não seja verdadeiro. Tal resulta do facto de nesses casos ser possível a produção de previsores, para cada elemento da amostra e/ou população, que contribuem para a redução da incerteza. Estas ideias foram exploradas por Särndal e Swenson (1987) e Bethlehem(1988), que sugerem o uso de informação auxiliar relevante quando ela existe. Os estimadores informativos sugeridos, deverão ser modificados de forma a contemplarem a modelização das não-respostas. Os estimadores informativos [ver Särndal (1992 - Cap. 9)], terão então a seguinte expressão[22]:

[22] Substituindo π_{kls} pelas estimativas obtidas, através da formação de grupos homogéneos de resposta.

$$\hat{t}_{c\pi^*} = \overbrace{\sum_{k\in U}\hat{y}_{1k}}^{\text{pré-estimativa}} + \sum_{h=1}^{H_s}\left[\overbrace{\sum_{k\in s_h}\frac{\hat{y}_k - \hat{y}_{1k}}{\pi_k}}^{\text{erro de pré-estimativa}} + \overbrace{f_h^{-1}\sum_{k\in r_h}\frac{y_k - \hat{y}_k}{\pi_k}}^{\text{erro de não-resposta}}\right]$$

com $\hat{y}_k$ e $\hat{y}_{1k}$, referidos à informação auxiliar da amostra e do Universo respectivamente

$\hat{y}_k = f_s(\mathbf{x}'_k)$ $\quad k \in s$ - forte previsor de y_k

$\mathbf{x}'_k$ - vector com n variáveis auxiliares

$\hat{y}_{1k} = f_U(\mathbf{x}'_{1k})$ $\quad k \in U$ - fraco previsor de y_k

$\mathbf{x}'_{1k}$ - vector com n_1 ($\leq$ n) variáveis auxiliares

que apresenta duas formas particulares, para os casos em que existe apenas informação auxiliar recolhida previamente à amostragem (isto é, relativa ao Universo) ou apenas informação auxiliar recolhida após a amostragem (isto é, relativa à Amostra):

Se apenas existir informação auxiliar a nível da amostra:

$$\hat{t}_{c\pi^*} = \sum_{h=1}^{H_s}\left[\sum_{k\in s_h}\frac{\hat{y}_k}{\pi_k} + f_h^{-1}\sum_{k\in r_h}\frac{y_k - \hat{y}_k}{\pi_k}\right]$$

Se apenas existir informação auxiliar a nível da população:

$$\hat{t}_{c\pi^*} = \sum_{k\in U}\hat{y}_{1k} + \sum_{h=1}^{H_s}\left[f_h^{-1}\sum_{k\in r_h}\frac{y_k - \hat{y}_{1k}}{\pi_k}\right]$$

7. Modelização do Mecanismo de Respostas Avaliação através da Simulação de Monte Carlo

(i) É importante voltar a referir, que se está a trabalhar com um modelo de respostas assumido que este poderá não coincidir com o verdadeiro mecanismo de resposta, o que na prática, significa que a modelização do mecanismo de resposta não elimina completamente o enviesamento. Estudos de simulação, usando o método de Monte Carlo, foram realizados por Särndal e Swenson (1987) e mostraram que o uso de mecanismos de resposta "não-exactamente" identificados, eram preferíveis a nenhuma modelização, quer em termos de enviesamento, quer ainda em termos da probabilidade de cobertura[23]. Por outro lado, mostraram ainda que as qualidades de não enviesamento e a probabilidade de cobertura melhoram se na fase de estimação se usar um estimador informativo.

(ii) A abordagem sugerida, estima as probabilidades de resposta admitindo homogeneidade de comportamentos no interior de cada grupo. Supõe-se, é claro que os grupos formados tem dimensões, que tornam possível a sua estimação. Uma diferente abordagem, consiste na estimação das probabilidades individuais de resposta (desiguais), com base numa abordagem paramétrica (Cassel, Särndal e Wretman 1983) ou não-paramétrica (Giommi 1988), usando informação auxiliar, correlacionada com a variável de interesse. A abordagem de Greenlees, Reece e Zieshang (1982) não só usa a variável de auxiliar, como também a própria variável de interesse. Seguindo uma sugestão de Gorieroux(1989) - embora com uma metodologia diferente

[23] Admitindo que o modelo possui alguma verdade.

Caldeira (1995) propôs o uso dos modelos GLM como forma de unificar as diferentes abordagens apresentadas.

(iii) Para realizar a simulação de Monte Carlo, Särndal e Swensson consideraram uma população de N=1227 famílias suecas, das quais havia informação relativa ao rendimento disponível (y) e ao seu rendimento fiscal (x) as quais apresentam um coeficiente de correlação de 83%. De acordo com o rendimento fiscal foram repartidos por 4 estratos com diferentes probabilidades de resposta. Em cada estrato os seus elementos tinham uma probabilidade de resposta (θ_h) uniforme (hipótese de trabalho para a simulação). A probabilidade de resposta apresentava, de acordo com o modelo, um coeficiente de correlação de 44% com o rendimento disponível. Apesar de baixa, esta correlação tem efeitos graves no enviesamento conforme adiante mostraremos. As características da população de referência são sintetizada no seguinte quadro:

Estratos	N_h	θ_h	$\bar{y}_h$	$\bar{x}_h$
U_1	373	0.45	4.18	1.68
U_2	303	0.65	5.65	2.25
U_3	280	0.75	6.73	2.54
U_4	271	0.90	8.31	2.79

(iv) Foi gerada uma amostra aleatória simples (SI) de 400 elementos da população das 1227 famílias. Esta amostra de 400 elementos, simulava responder ao inquérito aleatoriamente, de acordo com o modelo atrás descrito. Foram gerados assim 1000 conjuntos de respondentes, cada um deles a suportar uma estimativa, para um dado estimador.

(v) Foram considerados, para efeitos de estimação, três hipóteses de modelização do comportamento dos respondentes, saber:

1. Conhece-se o comportamento dos inquiridos, com diferenciação perfeita dos 4 estratos atrás referidos. Hipótese desejável mas irrealista, adiante designado como Modelo Perfeito.

2. Conhece-se estatisticamente o comportamento dos inquiridos, com diferenciação imperfeita em 2 estratos, $U_1 + U_2$ e $U_3 + U_4$ [24]. Hipótese mais realista, adiante designado como Modelo Imperfeito.

3. Desconhece-se o comportamento dos inquiridos e não se realiza esforço para a percepção de comportamentos diferenciados, isto é, supõem-se identicos. Hipótese pouco realista, adiante designado como Modelo Ausente.

(iv) Foram ainda considerados três estimadores $\hat{t}_A$, $\hat{t}_B$ e $\hat{t}_C$, o primeiro dos quais não usa informação auxiliar e os outros usam-na com base nos modelos de regressão.

(v) Com base nos estimadores apresentados e nas várias hipóteses de modelização do mecanismo de resposta, procedeu-se à apreciação estatística do enviesamento relativo, que é nulo se o estimador for não-enviesado.

[24] O sinal + tem o significado de reunião na Algebra de Conjuntos

Enviesamento Relativo Estimado (%): $B\left(\hat{t}\right) / t$

Modelo	$\hat{t}_A$	$\hat{t}_B$	$\hat{t}_C$
Perfeito	0%	0%	0%
Imperfeito	1.5%	0.3%	0.5%
Ausente	6.5%	1.7%	1.6%

(vi) Procedeu-se ainda à apreciação da taxa de cobertura dos Intervalos de Confiança Nominais a 95% , conforme consta do quadro. Esperava-se que 95% dos intervalos aleatórios estimados, por amostragem, contivessem o parâmetro a estimar. Contudo a simulação reservou-nos algumas surpresas.

Taxa de Cobertura do Intervalo de Confiança

Modelo	$\hat{t}_A$	$\hat{t}_B$	$\hat{t}_C$
Perfeito	95.2%	95.5%	95.6%
Imperfeito	93.3%	95.6%	95.6%
Ausente	46.3%	92.6%	93.9%

(vi) A **leitura dos quadros em coluna** revela que a modelização do mecanismo de resposta, reduz o enviesamento e aumenta a taxa de cobertura dos Intervalos de Confiança Nominais. Quanto mais perfeita for a modelização do mecanismo de resposta, maior será a qualidade das estimativas produzidas, como aliás era esperado. Mas o que é importante, uma vez que a modelização perfeita não é possível no tratamento realista das não-respostas, é o facto de **que a modelização mesmo imperfeita trás importantes**

ganhos de qualidade na estimação, qualquer que seja o estimador usado.

(vii) A **leitura dos quadro em linha** evidência uma redução do enviesamento e um aumento da taxa de cobertura se se fizer uso da informação auxiliar, no caso através do estimador da regressão. Em resultado, se **o uso de informação auxiliar relevante contribui para tornar robusta a estratégia de estimação** é obvio que melhor solução será obtida **combinando a modelização do mecanismo de resposta com o uso de informação auxiliar relevante**.

8 - Imputação de Respostas - Estratégia na Fase de Estimação

(i) A ideia de imputação faz sentido em dois casos essencialmente:

1. Se tivermos $s - r_c \neq \varnothing$, então faz-se a imputação para valores em falta e não se procede a qualquer modelização das respostas;
2. Se tivermos $r_u - r_c \neq \varnothing$, então faz-se a imputação para os valores em falta deste sub-conjunto e procede-se à modelização das respostas para $s - r_u$.

(ii) Embora seja possível fazer uma estimação específica para cada uma das questões, essa possibilidade levanta algumas questões importantes:

1. Em primeiro lugar, a cada variável de interesse pode estar associada um diferente modelo de tratamento de respostas, isto é, diferentes RHG, o que torna impraticável o uso destes modelos em inquéritos com muitas questões.

2. No caso de se pretender estimar o coeficiente de correlação, pode dar-se o caso de se encontrarem valores inadmissíveis para as estimativas (por exemplo: Coeficiente de correlação $\rho \notin [-1,1]$), uma vez que cada um dos totais que intervém na sua determinação assenta em diferentes conjuntos de respondentes e diferentes modelos de tratamentos de respostas. Este problema existe com frequência, quando se considera a estimação de parâmetros que envolvem mais do que uma variável de interesse.

A possibilidade mais óbvia é considerar para efeitos de estimação apenas os elementos que responderam a todas as perguntas do inquérito (r_c). Neste caso, haverá alguma perda de informação, o que pode até ser aceitável em muitos casos. Contudo, a solução mais sensata é proceder à imputação, de modo a ser possível a estimação, considerando todos os indivíduos que responderam a pelo menos uma questão do inquérito (r_u), isto é, sem perda de informação. Considera-se, nesta estratégia, que o custo de uma inadequada imputação é inferior ao da perda de informação.

(iii) A imputação é sempre uma operação delicada e, pelo menos do ponto vista estatístico, não se pode considerar aceitável colocar no mesmo plano valores observados e valores imputados. A questão da qualidade dessas imputações é muito importante e disso depende a qualidade da estimação. Só no caso de haver fundadas razões para considerar que estamos perante "bons" valores imputados é que poderemos considerar que a imputação é um bom método para o tratamento de respostas.

(iv) Uma boa **imputação, em geral, entra em consideração com informação auxiliar que permita imputar com alguma objectividade os valores pretendidos. O conhecimento das** propriedades estatísticas dos processos de imputação é um importante instrumento da sua avaliação. Estando em causa a qualidade dessa imputação, alguns países exigem através de legislação adequada que os ficheiros de dados individuais assinalem os valores sempre que estes resultem de uma imputação, indicando a sua origem artificial.

(v) Dos vários métodos de imputação usados, apresentam-se alguns, sem qualquer pretensão de exaustão.

Imputação Dedutiva - Neste caso, raro na prática, existe informação auxiliar sobre o valor em falta, que permite calcular com exactidão o valor em falta. A dedução pode inclusive basear-se nas outras variáveis disponibilizadas pelo inquérito.

Imputação Média - Os valores em falta de uma dada variável são imputados, com o valor médio dos valores respondidos. Embora, não se introduza enviesamento por esta via, deve dizer-se que este procedimento subestima a variância e o intervalo de confiança não é tecnicamente credível. Para minimizar este inconveniente, há quem advogue a modificação do estimador da variância.

Imputação Média da Classe - Usando informação auxiliar, a amostra é repartida de forma a que os grupos apresentem alguma homogeneidade. Os valores em falta são substituídos pela média dos respondentes da classe a que pertence o indivíduo que originou o valor

em falta. A vantagem deste método é que se houver algum desfasamento do valor imputado relativamente ao verdadeiro valor, espera-se que esse desfasamento seja menor do que no caso da imputação média. Tal como no caso anterior, também aqui se advoga a modificação do estimador da variância.

Imputação Aleatória (a quente[1]) - Com base nos valores observados pelo inquérito, escolhe-se aleatoriamente um de entre estes para imputar ao valor em falta. Embora o método pretenda atribuir variabilidade ao valor imputado, nada garante, pelo menos teoricamente, a qualidade da imputação.

Imputação Aleatória por Classes (a quente) - É evidentemente uma variante aperfeiçoada da imputação aleatória, quando existe informação auxiliar para se fazer a partição em classes homogéneas.

Imputação Sequencial a Quente[2] - Sempre que se encontra na presença de um valor em falta, percorre-se o ficheiro de dados para trás, até se encontrar um indivíduo da mesma classe (definida por variáveis auxiliares), que é o doador do valor em falta. É um procedimento não-aleatório, que apresenta a vantagem de, num único percurso, se realizarem todas as imputações. Note-se, no entanto, a possibilidade de numa única classe haver vários doadores. Para um bom funcionamento do procedimento aconselha-se a criação

[1] Expressão utilizada quando a imputação, se fundamenta nos valores observados pelo inquérito.

[2] O uso deste método é bastante frequente nos Estados Unidos, em Inquéritos Correntes à População.

de um número considerável de classes. O procedimento arranca com um doador por defeito para cada classe (imputação a frio[3]), em geral resultante de informação histórica.

Imputação através da função Distância (a quente) - É definida uma função distância que, usando a informação das restantes variáveis auxiliares disponíveis, permite identificar o respondente mais próximo. Ao valor em falta é atribuído o valor observado (da mesma variável) do respondente mais próximo.

Imputação por Regressão - É estimada uma regressão entre o valor da variável em falta para um dado indivíduo e os valores das variáveis disponíveis (além de outra informação auxiliar), com base nos valores fornecidos pelos respondentes à questão. A regressão estimada permite efectuar a imputação, com base na restante informação que o não-respondente a essa questão forneceu. Por vezes, acrescenta-se um resíduo, para se introduzir alguma aleatoriedade ao valor imputado.

Imputação Múltipla - Para cada valor em falta, realizam-se múltiplas imputações.
Calcula-se uma única estimativa (a média) com base nessas múltiplas imputações.
As maiores vantagens do método residem no facto de o estimador da variância (pooled) considerar não apenas a

[3] A imputação diz-se a frio se não tira partido da informação recolhida com o inquérito, embora possa fazer uso de informação não fornecida pelo inquérito.

incerteza da estimativa mas também a incerteza da imputação.

(vi) Apesar de ser uma técnica bastante popular, não existe literatura abundante sobre os métodos de imputação, as suas propriedades e/ou estudos empíricos sobre as consequências dos procedimentos. Para a popularidade da técnica contribuiu sobretudo a simplicidade das suas versões mais elementares e o crescente uso da amostragem acompanhada de não-respostas, ao longo das últimas décadas.

(vii) Sem tecer considerações definitivas, a perspectiva do texto sugere o uso da imputação como técnica auxiliar destinada a contornar o peso excessivo de algumas abordagens sugeridas. Por essa razão, a imputação na presente abordagem tem um papel meramente acessório.

CAPÍTULO VII - TÓPICOS SOBRE SONDAGENS ELEITORAIS

1 - Generalidades

Nas sondagens de opinião, a avaliação dos resultados supõe implicitamente a consideração de uma **re-análise interna** e de uma **re-análise externa**.

A primeira está relacionada com a técnica de amostragem, de um modo geral testada no âmbito das ciências sociais e humanas, sem prejuízo do importante contributo que os estatísticos poderão dar na análise das implicações resultantes das múltiplas opções tomadas quando se elabora uma operação estatística.

Relativamente à **re-análise externa**, uma sondagem é, de facto, um somatório de opiniões e intervenções individuais que emergem no quadro complexo de tensões sociais e de novos paradigmas em matéria de acesso à informação o que significa que os resultados da sondagem de nada valerão se não forem convenientemente analisados e comentados.

Focaremos agora o caso particular das sondagens eleitorais as quais desempenham um papel insubstituível em sociedades como a nossa, exteriorizando desde logo um direito próprio da vivência democrática - o direito à opinião, o direito que cabe a todo o cidadão de manifestar a sua discordância ou concordância por esta ou aquela política exprimindo quotidianamente o seu descontentamento ou a sua incompreensão por eventuais iniciativas governamentais e contribuindo para o esboço de uma opinião colectiva repleta de mensagens particularmente úteis aos decisores políticos.

As sondagens eleitorais começaram a ter algum impacto junto da opinião pública nos Estados Unidos dos anos 30 a propósito das projecções eleitorais realizadas nas vésperas das eleições Presidenciais de 1936 (como já foi referido na nossa introdução geral). Assim assistiu-se logo na década seguinte ao seu desenvolvimento na Europa e muito especialmente no Pós-Guerra a recorrência às sondagens eleitorais tornou-se uma prática comum em todos os Países da Europa Ocidental munidos de Regimes Democráticos.

Note-se que nas sondagens eleitorais é necessário compreender que o **voto** propriamente dito é um acto, isto é, **um comportamento** enquanto que a sondagem **regista apenas uma intenção.**

Assim as sondagens eleitorais revelam a **tendência do voto no momento de inquirição** e **não podem** ser confundidas com as previsões para o dia das eleições; estas últimas podem basear-se parcialmente no resultados de diversas sondagens consecutivas mas assentam fundamentalmente em **modelos** que se alicerçam no comportamento eleitoral em actos eleitorais anteriores com vista a realizar uma previsão da distribuição dos **indecisos e não respondentes.**

Em todo o caso, apesar da ênfase que os orgãos de Comunicação Social acabam por dar às sondagem eleitorais que antecedem o acto eleitoral, não é geralmente explicitado se se trata de uma última sondagem ou de uma previsão eleitoral e neste último caso raramente são explicitados os argumentos que alicerçam a modelização dos indecisos. Esta situação admite adicionalmente diferentes matizes nos Países da Europa Ocidental em termos da Lei eleitoral vigente que autoriza nuns casos a divulgação dos resultados das

sondagens até à data das eleições e noutros não autoriza tal divulgação na vizinhança temporal do acto eleitoral. Esta questão não pode deixar de ter algum impacto já que em Países como Portugal (ver anexo 5) onde não é permitido divulgar os resultados nos sete dias que antecedem as eleições, as últimas sondagens realizadas pelas diferentes empresas do sector são de um modo geral divulgadas no fim de semana anterior ao acto eleitoral, isto é, tais sondagens exprimem a intenção do eleitorado cerca de quinze dias antes do momento eleitoral. Ora este facto vem revelar que em Países como o nosso os resultados das sondagens Pré-eleitorais não podem tomar ainda em linha de conta o **efeito da Campanha** eleitoral, o que constitui um ponto fraco do modelo previsional, caso exista.

Apesar destas considerações preliminares, o balanço das sondagens eleitorais nos vários Países do Mundo onde elas se realizam (incluindo Portugal) não pode deixar de ser extremamente positivo se tomarmos em consideração o confronto entre as previsões induzidas directa ou indirectamente pelas sondagens e os resultados propriamente ditos no dia do acto eleitoral. Contudo a experiência destas décadas mostram à evidência que as sondagens também são falíveis por várias razões que tentaremos explicar; assim o mostra os insucessos ocorridos em diferentes Países do Mundo, dos quais destacaremos os seguintes:

- em Espanha no ano de 1993 as várias sondagens levadas a cabo revelaram uma **tendência do eleitorado** conducente à vitória do PP, isto é, à queda do Governo Socialista de Felipe Gonzalez, em boa parte na sequência do "Assunto Filesa" que gerava suspeições quanto a possíveis corrupções no seio desse Governo. Mesmo na sondagem à boca das urnas uma das sondagens continuava a estimar a vitória do PP e as outras duas a vitória do PSOE por uma

escassa margem. A verdade é que os resultados finais deram uma vitória nítida ao PSOE que arrecadou 159 deputados contra 141 para o PP. Tais discrepâncias geraram uma avalanche de críticas e indignações as quais desacreditaram o sector.

Algo muito semelhante já tinha ocorrido em 1948, nos Estados Unidos conhecido por "**Erro Gallup**" onde o Instituto Gallup (célebre neste País) previa a vitória do Candidato Republicano Thomas Dewey sobre o Candidato Democrata Harry Truman (vice-presidente de Roosevelt que assumiu a presidência após a morte daquele) e o resultado final propiciar a vitória a TRUMAN com avanço de 4 pontos percentuais.

Um terceiro fracasso que foi objecto de muito eco na opinião pública ocorreu em Inglaterra na Primavera de 1992, poucos meses após a demissão de Margaret Tharcher: John Major convoca eleições antecipadas e as sondagens convergiram todas num cenário de vitória dos trabalhadores; mesmo duas das sondagens à boca das urnas, encomendadas por cadeias de televisão, apontavam para a vitória dos trabalhistas. Tal não veio a ocorrer por Major vence confortavelmente. Esta situação abalou o País mais liberal e flexível da Europa quanto à legislação sobre a difusão das sondagens.

Poderíamos ainda referir com destaque os fracassos das eleições Presidenciais Francesas em 1995 ou das eleições legislativas de Itália em 1994 (consultar por ex. West J., 1996).

Para além de alguns aspectos já referidos neste texto e que podem ser explicativos de situações de insucesso tentaremos trazer mais algumas considerações que no fundo exemplificam a complexa teia subjacente à "captura" deste tipo de conhecimento.

Ao nível da sondagem eleitoral propriamente dita e no que se relaciona com o processamento de dados confrontam-se essencialmente duas abordagens:

- **A ponderação** com base nos resultados em votos anteriores em eleições similares, isto é, conhecemos a distribuição dos votos na eleição anterior e relativamente aos inquiridos pergunta-se em que partido vão votar e em que partido votaram nas eleições anteriores.

- **A não ponderação** por resultados eleitorais

Observe-se que não fazer este tipo de ponderação suscita desde logo grandes dúvidas quanto à "representatividade" da amostra no que diz respeito à estrutura partidária.
O recurso a outras variáveis auxiliares tais como o sexo e o escalão etário não assegura esse tipo de representatividade; e por outro lado a selecção de "pontos de amostragem" mesmo que derive de um desenho amostral sensível ao espectro político do País, não reflectirá com suficiente rigor a estrutura presente ou passada do voto. Note-se que a pós-estratificação implícita na reponderação não irá alterar sensivelmente as estimativas pois se tal ocorresse isso significaria que o desenho amostral que originou a amostra não era aceitável.

Os opositores à ponderação pelos resultados eleitorais das eleições passadas criticam esta abordagem argumentando que os entrevistados terão a tendência para afirmar que votaram no passado no partido que pensam votar agora, eventualmente por alguma retracção em revelar aos entrevistadores que mudaram (que se enganaram). Acresce ainda outra crítica segundo a qual se afirma que os respondentes terão esquecido

em muitos casos a sua escolha de voto nas eleições anteriores.

Reconheça-se que em toda esta polémica há situações paradoxais: por exemplo, em vários inquéritos visando a taxa de fidelidade de voto, não se tem constatado, muito pelo contrário, elevadas taxas de fidelidade: os resultados apontam para os quatro maiores partidos de Portugal, taxas de fidelidade entre 50% e 80%. Complementarmente todos sabemos que a sociedade actual é bem mais volátil no que diz respeito a preferências partidárias, por várias razões de cariz sociológica e decorrentes do enquadramento geo-político de Portugal na Europa e no Mundo. Ora estudos realizados em vários países revelam que a probabilidade de indicar o voto nas eleições em jogo não é sempre igual porque depende das tendências partidárias do inquirido o que pode constituir mais um argumento a favor da reequilibragem da amostra. Em todo o caso este problema justificaria uma análise estatística profunda que permitisse balancear as duas abordagens através dos seus pontos fortes e fracos e sobretudo encontrar modelos mais sofisticados relativos ao mecanismo de resposta, certa ou errada, por parte do inquirido, isto é, variáveis que permitam discriminar os indivíduos que fazem **o jogo da verdade** relativamente à questão do voto anterior e aqueles que fazem **o jogo da mentira**.

Complementarmente, a interrogação sobre o voto anterior não pode nem deve surgir no questionário numa posição arbitrária mas sim em função da posição da pergunta relativa ao voto actual: convém que as perguntas estejam suficientemente distanciadas.

Os Institutos de sondagens realizam de um modo geral as chamadas sondagens "**à boca das urnas**" (**exit polls**) as quais são efectivadas ao longo do dia das eleições na vizinhança de uma amostra representativa de assembleias de voto inquirindo uma amostra aleatória de indivíduos que acabaram de votar: pretende-se conhecer o sentido de voto e respectivas razões a par de características socio-económicas do eleitor; em princípio é o próprio eleitor que preenche o mini-questionário e as respostas vão sendo transmitidas para o centro de apuramento de dados garantido a divulgação dos resultados logo após o encerramento das urnas. Trata-se pois de uma sondagem pós-eleitoral.
Naturalmente que estas projecções alicerçam a primeira parte dos debates televisivos na "longa" noite de escrutínio dos resultados finais.

Por vezes também se recorre a um inquérito via telefone pouco antes do encerramento das urnas, junto de uma amostra "representativa" de eleitores o que permite obter também informações sobre a incidência de abstenções e respectiva interpretação.
Relativamente ao "exit poll" a teoria das sondagens intervêm na selecção dos centros de votação utilizando como critérios de estratificação, a Região e o habitat (dimensão do círculo eleitoral).
A alternativa poderia ser o chamado critério de afectação óptima baseado na observação dos resultados das eleições anteriores nas distintas secções de voto de cada círculo eleitoral e a inclusão na amostra daqueles que apresentam menores desvios relativamente ao resultado nacional. Este método pressupõe uma grande estabilidade na composição das secções eleitorais em diferentes momentos eleitorais o que nem sempre se verifica (Espanha é um País onde este método é por isso pouco recomendado).

Quanto à selecção das pessoas toma-se como critério inquirir uma em cada X pessoas que saem do local de voto (com eventual alteração do passo nos períodos do dia de maior afluência às urnas). Por fim a sondagem "à boca das urnas" é complementada frequentemente por uma outra (**quick count**) segundo a qual os mesmos entrevistadores que ao logo do dia entrevistaram os votantes, observam, após o fecho das urnas, o escrutínio nalgumas mesas de voto dessa secção de tal forma que os resultados obtidos corrigem a primeira projecção realizada.

Esta discussão, sempre rica, ocorre normalmente porque a opinião pública incorre sistematicamente no erro que é o de confundir sondagens com previsões eleitorais.

E quem é que não tem uma curiosidade de comparar os resultados das sondagens com os resultados efectivamente verificados nas eleições?

Para ilustrar tudo isto, vejamos quadros comparativos dos resultados de sondagens apresentados pelos vários institutos em Portugal, nas eleições legislativas de 1991 e de 1995; e analisemos ainda a sua distância ao verdadeiro *score* obtido por cada um dos partidos no acto eleitoral.

ELEIÇÕES LEGISLATIVAS 1991

Data	Título	Instituto	PSD	PS	CDU	CDS
21/9	Expresso	Euroexpansão	44.2	35.8	7.9	5.2
27/9	DNotícias	Marktest	45.2	34.5	8.1	4.8
27/9	Jornal	Pluriteste	41.2	34.7	8.4	8.1
28/9	Público	Euroteste	47.3	35.5	8.5	4.1
28/9	Expresso	Euroexpansão	44.2	33.1	9.4	6.2
28/9	JNotícias	Euroteste	46.4	37.0	9.7	4.1
Resultado eleições			*50.4*	*29.3*	*8.8*	*4.4*

ELEIÇÕES LEGISLATIVAS 1995

Data	Título	Instituto	PSD	PS	CDU	CDS
21/9	Visão	Marktest	31	49	6	7
22/9	Independente	Metris	29	34	9	8
23/9	Público	UCatólica	35	40	9	10
23/9	Expresso	Euroexpansão	32.7	44.5	11.5	6.9
23/9	Semanário	Compta	40	41	10	7.5
23/9	Dnotícias	Euroteste	35	39	9	10
23/9	JNotícias	Ipsos	29	34	8	8.5
Resultado eleições			*34.1*	*43.9*	*8.6*	*9.1*

Em qualquer dos casos analisados, podemos verificar uma relativa proximidade entre os resultados apresentados pelos vários Institutos; bem como se constata que em ambos os casos, o partido efectivamente vencedor das eleições era considerado como tal nas últimas sondagens realizadas em cada um dos períodos eleitorais.

A dimensão destas discrepâncias pode ser entendida como um índice de qualidade; com efeito, ao tentar implementar um bom desenho do inquérito, um instituto pode ser mais bem sucedido do que um seu concorrente. Mas é também certo que um mesmo resultado de trabalho de campo pode dar origem a resultados diversos, se diversa for a estratégia de sondagem utilizada. Há mesmo, nos casos acima referidos, situações em que os mesmos resultados de campo originam diferentes resultados finais. Essas diferenças resultam da utilização ou não das técnicas de tratamento de indecisos e da ponderação (ou não) política da amostra (utilizadas estas técnicas de sondagem em conjunto ou separadamente).

Consideremos agora um outro importante aspecto inerente às sondagens eleitorais e que se relaciona com **a possível influência que os resultados dessas sondagens podem exercer sobre o eleitorado** - este tema tem sido objecto de grandes controvérsias pelo facto de estar directamente relacionado com a maior ou menor flexibilidade das leis eleitorais em matéria de divulgação dos resultados da sondagem.

Aqueles que consideram que tais resultados podem manipular a opinião pública tendem a impor maiores limitações ao processo de divulgação argumentando que tal influência sobre a opinião peca por não ser revertível já que aqueles que são influenciados, isto é, os eleitores, não têm a mesma possibilidade de influenciar aqueles que, nesse instante, os influenciam, ou seja, falha um princípio chave das Democracias: a simetria. Ora este tipo de considerações podem ser criticáveis se nos posicionarmos nas sociedades abertas em que vivemos onde as sondagens eleitorais são frequentemente indutoras de processos de mudança.

Complementarmente será possivelmente **redutor** partirmos da ideia de que alguém tem melhor conhecimento daquilo que é melhor para o eleitor, **quando a democracia se baseia no princípio segundo o qual cabe ao eleitor decidir**!

É neste contexto que a lei Portuguesa actual (ver em anexo a lei nº 31/91) proíbe a divulgação de resultados na semana anterior às eleições, decisão criticável por muitos representantes deste sector, enquanto noutros países essa limitação é circunscrita aos dois dias anteriores e noutros não há qualquer tipo de restrição.

Finalmente refira-se um aspecto importante previstos na actual lei 31/91 sobre a Publicação e difusão das sondagens e inquéritos de opinião e que consiste na obrigatoriedade, a que está sujeita a empresa responsável pela publicação ou difusão

de sondagens, em depositar, junto da Alta Autoridade para a Comunicação Social, um rol de informações - **a chamada ficha técnica** - até ao dia da difusão dos resultados.
Complementarmente a lei estabelece quais são os elementos dessa ficha técnica que devem acompanhar toda e qualquer difusão dos resultados de sondagem eleitoral - infelizmente na generalidade dos Países Europeus é não só frequente a não observância destes princípios, já que a ficha técnica vulgarmente difundida, é demasiadamente pouco informativa, como se assiste a vários atropelos quanto à forma como são divulgados os resultados de tais sondagens, dando-se por vezes lugar de destaque a informações estatísticas não merecedoras de tais adjectivações!
E neste campo há também alguma polémica entre aqueles que consideram que o mercado acabará por fazer a selecção natural - entre os orgãos de comunicação que informam de facto o público e os que para tal não são capazes e outros que não aceitam tais argumentos em mercados reduzidos como o nosso onde a "lei do mais forte e mais capaz" ainda não se fez sentir em boa parte devido à relativa indulgência do mercado quando jornais de grande prestígio divulgam uma sondagem que não merece a mínima credibilidade. Ou seja, a credibilidade do sector é globalmente posta em causa (devida à dimensão global do sector) aconselhando-se uma maior vigilância por parte da Alta Autoridade para a Comunicação Social na detecção de tais situações , nos termos da lei 31/91. (Consultá-la em anexo).

Finalmente abordaremos o problema do chamado **televoto** recorrendo para o efeito a parte de um artigo publicado na "Vida Económica" no corrente ano e assinado por Paulo Morais:

Televoto ou telelixo?

O advento em Portugal das televisões privadas trouxe alterações profundas na relação entre os canais de TV, enquanto emissores, e os seus destinatários ou receptores, isto é, os telespectadores. Alterações decorrentes de um novo tipo de informação, mais agressiva, com o que isso comporta de positivo, mas também de perverso; modificações resultantes da proliferação de programas mais vulgarmente designados de talk-shows, fazendo um apelo ao que de mais íntimo há em cada um de nós e tirando proveito das nossas vulnerabilidades individuais e da fragilidade da nossa estrutura social; e muitas outras inovações, novos actores, revelação de talentos até então escondidos; e, claro, o surgimento e difusão maciça da interactividade em televisão, através de concursos, passatempos, telefonemas de valor acrescentado. Ocorreu uma completa metamorfose no panorama do audiovisual; uma revolução mesmo. Revolução com muitas virtudes mas também, como todas as revoluções aliás, com muitos atropelos ao bom senso e à sabedoria, ao sentido ético, e até à legalidade.

Nos modelos adoptados de interactividade em televisão, tomou particular predominância esse tão popular passatempo que é o televoto, espécie que parece ter vindo para ficar, para mal dos nossos pecados.

O vulgarmente designado televoto, não constituindo uma sondagem, nem tão pouco uma auscultação de opinião, é percepcionado pelo telespectador como se tratando de uma

sondagem; aliás, a isso o induz a palavra voto que comporta em si mesmo a carga conceptual de consulta.
E não constitui uma sondagem, em primeiro lugar porque não está cumprindo o requisito mínimo em qualquer consulta, que é o de termos definido o universo em estudo.
Acrescente-se ainda que quando a questão é repetida sucessivas vezes, há uma alteração desse hipotético universo. Não estando definido o universo, não será pois legítimo falar em sondagens, nem tão pouco induzir os telespectadores a pensar que assim é.
Em segundo lugar, e imaginando que a credibilidade da iniciativa resistiria ao argumento anterior, perante a existência dum universo virtual, haveria que encontrar critérios de amostragem (preferencialmente aleatórios ou probabilísticos). Não temos aqui uma amostra (probabilística ou não), uma vez que as respostas obtidas não obedecem a qualquer critério de amostragem. Nem sequer a qualquer tipo de critério. Os eventuais respondentes serão os telespectadores mais motivados, ou os menos ocupados; mas seguramente os mais ingénuos. Sendo que, para além do mais, um telespectador pode responder mais que uma vez, contribuindo assim para o enviesamento da alegada amostra. Como se tal não bastasse, não há sequer tempo para que possam responder todos aqueles que o pretendam, uma vez que o tempo disponibilizado entre a colocação da questão e o termo do passatempo é escasso (mesmo quando a questão é colocada em dias sucessivos, tal situação diz respeito à última colocação da questão); não podem pois participar todos aqueles que o pretendam. Nestas circunstâncias, não se torna possível descrever a hipotética amostra e o perfil dos indivíduos que a constituem, o que é primordial em qualquer trabalho de amostragem.

2 - Reforma do Sistema Eleitoral Português

2.1 - Introdução

Um sistema eleitoral, entendido, em sentido restrito, como modo de conversão dos votos em mandatos, é apenas um factor de apreciação do quadro em que a representação se define; não pode ser dissociado de muitos outros, como sejam a natureza do sufrágio, a dimensão dos círculos, a capacidade eleitoral, passiva e activa, as condições de propositura das candidaturas, o modo como são reguladas as campanhas eleitorais e o seu financiamento, e assegurada ou não, melhor ou pior, a igualdade entre os candidatos, etc.

O sufrágio universal não é necessariamente contraditório com a imposição de restrições, ditadas por razões circunstanciais, como sejam a residência ou exercício de determinadas funções. Mas exclui todas aquelas que tenham por base o nível de rendimento, o grau de instrução, o sexo, a raça, a língua ou as convicções religiosas.

Há centenas de sistemas eleitorais actualmente em vigor e também inúmeras combinações entre alguns deles. De qualquer forma, e por uma questão de uniformização, podemos dividi-los em três famílias, se atendermos aos métodos de escrutínio: sistemas maioritários ou pluralitários, semi-proporcionais e proporcionais. Dentro destes, temos nove subfamílias. Os maioritários podem dividir-se segundo a terminologia mais comum, em *First past the Post (FPTP)*, *Block Vote (BV)*, *Alternative Vote (AV)* e *Two Round System (TRS)*, como sistemas maioritários. São sistemas semi-proporcionais os designados por *Parallel* e *Single non transferable vote (SNTV)*. Finalmente, *List Proportional Representation*, *Mixed Member Proportional (MMP)* e *Single Transferable Vote (STV)*, são os sistemas proporcionais mais comummente utilizados.(Anexo 6).

Enquanto nos métodos maioritários, é normalmente eleito ou eleitos os candidatos da lista ou listas mais votadas, já nos métodos proporcionais há que definir critérios (de proporcionalidade para distribuição de mandatos pelas diversas listas, em função, obviamente, da votação de cada lista.
Os métodos proporcionais pressupõem a atribuição de mandatos segundo um determinado critério. Os métodos mais comuns são os da utilização da quota eleitoral (proporcionalidade pura), métodos semi-proporcionais de Hondt, de Imperiali, de St. Lague ou St. Lague modificado.

2.2 - O Sistema eleitoral português

O sistema eleitoral português experimenta hoje, volvido que é um quarto de século de vida da democracia, profundas alterações.
A reforma do sistema eleitoral, a redefinição dos seus principais elementos estruturantes (modo de escrutínio e círculos eleitorais); uma actualização global do recenseamento, através da constituição de uma base de dados eleitoral centralizada e devidamente estruturada; a introdução e generalização de novas tecnologias nos processos e na logística eleitoral, nomeadamente com a utilização de recursos como o voto electrónico – todos estes elementos constituirão sem dúvida factores de profunda transformação no processo eleitoral português; e talvez até na forma de funcionamento do regime democrático.

Desenvolveremos de seguida dois aspectos fulcrais das referida metamorfoses que são os da reforma eleitoral e o da actualização do recenseamento:

2.2.1 - A reforma do sistema eleitoral português: comparação entre o sistema actual e as propostas dos dois maiores partidos

Situação actual (definida pela Lei n.º 14/79, de 16 de Maio – Lei Eleitoral para a Assembleia da República, LEAR)

Os círculos eleitorais do continente coincidem com as áreas dos distritos administrativos, são designados pelo mesmo nome e têm como sede as suas capitais. Acrescem os círculos das regiões Autónomas da Madeira e dos Açores e os círculos da Europa e Fora da Europa.
Número total de deputados é de 230.
O número de deputados pelos círculos eleitorais do território nacional é de 226, distribuídos proporcionalmente ao número de eleitores de cada círculo, segundo o método da média mais alta de Hondt (artigo 13.º da LEAR).
O critério de eleição é o de distribuição dos mandatos através da utilização do método de escrutínio de Hondt.
A conversão de votos em mandatos faz-se de acordo com o método de representação proporcional de Hondt, obedecendo às seguintes regras:

a) apura-se em separado o número de votos recebidos por cada lista no círculo eleitoral respectivo;
b) o número de votos apurados por cada lista é dividido, sucessivamente, por 1, 2, 3, 4, etc., sendo os quocientes alinhados pela ordem decrescente da sua grandeza numa série de tantos termos quantos os mandatos atribuídos ao círculo eleitoral respectivo;

c) os mandatos pertencem às listas a que correspondem os termos da série estabelecida pela regra anterior, recebendo cada uma das listas tantos mandatos quantos os seus termos da série;
d) em caso da subsistência de empate nos quocientes apurados, os mandatos são atribuídos aos partidos menos votados (esta é uma correcção ao método de Hondt).

Proposta do Partido Socialista

O número total de deputados deverá ser 230.
O número de deputados do território nacional é de 226, sendo 35 atribuídos ao círculo nacional e os demais distribuídos proporcionalmente ao número de eleitores de cada círculo eleitoral parcial, nos termos dos critérios abaixo definidos:
a) obtém-se a quota que resulta da divisão do número total de eleitores pelo número de deputados a distribuir;
b) o número total de eleitores recenseados em cada círculo parcial é dividido por essa quota;
c) a parte inteira do quociente assim obtido é logo convertida em lugares atribuídos ao círculo;
d) os restantes lugares são atribuídos por ordem decrescente aos círculos cujo quociente apresente um maior resto;

A proposta do Partido Socialista introduz círculos uninominais de candidatura.
A determinação destes é feita da seguinte forma: cada um dos círculos eleitorais parciais do continente é dividido em círculos uninominais em número igual a metade mais um dos lugares atribuídos ao respectivo círculo, arredondado por defeito se necessário.

O número de eleitores de cada círculo uninominal contem-se nos limites definidos pelos factores multiplicativos 0,8 e 1,2 do resultado da divisão do número de eleitores do círculo parcial pelo número de círculos uninominais em que este é dividido.
Em cada círculo parcial, o número de eleitores dos círculos uninominais deve apresentar a menor variância possível.
A conversão de votos em mandatos nos círculos eleitorais parciais faz-se de acordo com o método de Hondt. No círculo eleitoral nacional, a conversão de votos em mandatos faz-se pelo método de Hondt, aplicado ao resultado da soma das votações obtidas pelas listas nos círculos parciais do território nacional a que concorrem.
Dentro de cada lista concorrente a um círculo eleitoral parcial do continente os mandatos apurados são conferidos pela seguinte ordem:

a) em primeiro lugar, aos candidatos que tenham sido os mais votados de entre os candidatos concorrentes no respectivo círculo uninominal de candidatura;
b) a seguir, se ainda houver mandatos a conferir, . aos candidatos da lista plurinominal pela ordem de precedência da lista
c) se o número de candidatos nas listas uninominais for superior ao número de mandatos atribuídos no respectivo círculo eleitoral parcial à lista em que concorrem, o mandato é-lhes conferido, sendo subtraídos, em igual número, os mandatos que tiverem resultado para a mesma no círculo nacional.

Proposta do Partido Social Democrata

Cada eleitor dispõe de um voto no círculo da sua área territorial de recenseamento e de um voto no círculo nacional.

Os eleitores recenseados no território nacional distribuem-se por oitenta e cinco círculos uninominais, cobrindo todo o território continental, e por dois círculos regionais, cobrindo cada uma das Regiões autónomas dos Açores e Madeira.
Os eleitores residentes fora do território nacional são agrupados em três círculos das comunidades portuguesas, um abrangendo o território de Macau e os países de língua oficial portuguesa, outro os países europeus e o terceiro os demais países.
Há duplo voto. A cada um dos votos definidos corresponde um boletim de voto.
O número total de mandatos é de 184.
A distribuição dos círculos uninominais faz-se na proporção directa do número de eleitores recenseados em cada círculo.
A variação do número de eleitores entre círculos uninominais deve respeitar em cada distrito o intervalo de mais ou menos 1/3 do respectivo número médio de eleitores.
A conversão dos votos locais em mandatos faz-se:

a) nos círculos uninominais, atribuindo o mandato ao candidato mais votado;
b) nos círculos regionais e nos círculos das comunidades portuguesas, atribuindo os mandatos de acordo com o método de representação proporcional de Hondt.

A conversão dos votos nacionais em mandatos faz-se nos termos seguintes:

a) apura-se, relativamente ao número total de deputados, os mandatos que correspondam a cada candidatura, de acordo com o método de representação proporcional de Hondt
b) atribui-se a cada candidatura o número de mandatos igual à diferença entre o número apurado, nos termos da alínea anterior, e o número dos deputados por ela eleitos nos

círculos uninominais, regionais e das comunidades portuguesas

Se alguma das candidaturas eleger nos círculos uninominais, regionais e das comunidades portuguesas deputados em número superior ao que lhe caberia pelo apuramento referido em a), serão atribuídos às outras candidaturas tantos mandatos quantos os necessários para assegurar a proporcionalidade adequada.

Apreciação global das mudanças propostas

Em termos gerais, quer a proposta do Partido Socialista, quer a do Partido Social Democrata introduzem como característica os círculos uninominais, na perspectiva da aproximação entre eleitos e eleitores – mas sem que o sistema eleitoral perca as características de proporcionalidade que lhe são impostas pelo artigo 288.º da Constituição.
Sabe-se, no entanto e tanto quanto reza a ciência política, que são os sistemas maioritários que andam associados a círculos uninominais; enquanto que os sistemas proporcionais se casam com os tradicionais círculos regionais ou distritais. Acresce que os critérios que presidam à definição dos círculos eleitorais, as características e dimensão podem determinar um maior ou menor respeito pelo princípio de proporcionalidade constitucionalmente consignado; cabe pois aqui um papel fundamental aos juristas, mas também a sociólogos e matemáticos.
Esta é aliás uma discussão que haveria que fazer na sociedade portuguesa que é a de saber se a opinião pública portuguesa é efectivamente favorável a um sistema proporcional, tal como o impõe a Constituição ou alternativamente seria mais favorável a um sistema maioritário. Ou dito por outras palavras, há que saber se admitimos (ou não) que possa

existir uma maioria parlamentar, sem que esta necessariamente assente numa maioria eleitoral. Se o permitirmos, então poderemos optar por sistemas ditos maioritários, de círculos uninominais como o francês ou o inglês; se não, teremos de continuar com sistemas proporcionais puros ou enviezados, como é o caso do actual, em que pontifica o nosso já tão familiar método de Hondt.
No primeiro caso, tal obrigaria a uma alteração da Constituição, no articulado do seu artigo 288°. Mas, e segundo defendem alguns constitucionalistas, a alteração do artigo 288, que define os limites da revisão constitucional, obrigaria à elaboração de uma nova Constituição.

2.2.2 - Necessidade de actualização dos cadernos eleitorais. Processo de actualização; suas eventuais consequências nos círculos de apuramento e na geografia eleitoral portuguesa.

O processo de actualização dos cadernos eleitorais em Portugal, cuja primeira fase foi concluída em finais de Maio de 1998, surge na sequência da constatação da existência de discrepâncias entre os diferentes números de potenciais eleitores (número de recenseados na base de dados eleitoral do Ministério da administração Interna, estimativas do Instituto Nacional de Estatística para a população com mais de 18 anos e residente em Portugal, estimativas equivalentes das Nações Unidas,...).

A primeira vez que esta questão é levantada com acuidade é aquando do censo de 1991, levado a cabo pelo INE; constata-se então haver uma discrepância de aproximadamente 600000 indivíduos. Haveria pois que avaliar em que medida este fenómeno se iria corrigir com o evoluir dos anos ou se, pelo contrário, se agravaria esta diferença.

Em que medida são diferentes os números apresentados pelos cadernos eleitorais e os censos do INE, relativamente à população com idade de votar?

Para responder a esta questão, optemos por utilizar números relativos a 1995, uma vez que neste ano, para além de dispormos das estimativas de população para o território do continente e regiões autónomas que o Instituto Nacional de Estatística apresenta anualmente, decorreram eleições para a Assembleia da República.

Podemos pois, neste contexto, simular as consequências (em termos de representação geográfica e partidária) duma actualização do recenseamento, utilizando para tanto quer as estimativas de população em idade de votar, quer os resultados efectivamente obtidos nas eleições.

Numa primeira análise, e por simples comparação dos cadernos eleitorais de 1995 com as estimativas do INE do mesmo ano, relativamente à população com mais de dezoito anos - constata-se uma diferença de cerca de um milhão de indivíduos (O Quadro I apresenta essas diferenças).

QUADRO I

DISTRITO	REC. 95	ESTIM. INE 95	DIFERENÇA	DIFERENÇA RELATIVA
AVEIRO	559.718	513.990	45.728	8,17%
BEJA	152.031	129.880	22.151	14,57%
BRAGA	635.169	565.330	69.839	11,00%
BRAGANÇA	151.845	120.940	30.905	20,35%
C. BRANCO	203.039	168.940	34.099	16,79%
COIMBRA	382.941	339.340	43.601	11,39%
ÉVORA	151.251	137.270	13.981	9,24%
FARO	308.767	274.860	33.907	10,98%
GUARDA	175.602	145.910	29.692	16,91%
LEIRIA	374.841	337.170	37.671	10,05%
LISBOA	1.874.449	1.644.320	230.129	12,28%
PORTALEGRE	116.764	104.840	11.924	10,21%
PORTO	1.404.949	1.276.960	127.989	9,11%
SANTARÉM	395.852	354.760	41.092	10,38%
SETÚBAL	639.142	576.210	62.932	9,85%
V. CASTELO	225.243	194.090	31.153	13,83%
V. REAL	221.292	179.090	42.202	19,07%
VISEU	352.553	306.430	46.123	13,08%
AÇORES	188.868	170.340	18.528	9,81%
MADEIRA	205.088	188.820	16.268	7,93%
TOTAL	**8.719.404**	**7.729.490**	**989.914**	**11,35%**

Esta variação de cerca de um milhão de eleitores a mais reflecte-se naturalmente em todos os círculos, mas terá maior incidência nos distritos do interior (Bragança, Castelo Branco, Guarda e Vila Real).

Optamos neste estudo por considerar as estimativas do INE, até pelo facto de podermos trabalhar com dados geograficamente mais finos, nomeadamente ao nível do distrito. Esta variação seria tanto maior se os dados utilizados fossem os do Instituto IDEA (Institute for Democracy and Electoral Assistance), que considera para o nosso país como número de recenseados e de população em idade de votar os números de aproximadamente 8,9 e 7,5 milhões, respectivamente, o que daria uma diferença de cerca de um milhão e meio (vide Quadro II). Contudo se admitirmos que em 1991 o INE obteve uma taxa de cobertura de 97% poderíamos diminuir em cerca de 200 000 a referida discrepância, isto é, as informações disponíveis apontam como mais provável a existência de cerca de 800 000 situações irregulares que se vem acumulando nas últimas duas décadas.

QUADRO II

Ano	Total de Votos	Recenseados	Votos/Recen-seados	Pop.+18 anos	Votos/Pop. +18 anos	Pop. Nº de Hab.
1975	5 666 696	6 177 698	91.70%	6 315 420	89.70%	9 426 000
1976	5 393 853	6 477 619	83.30%	6 476 220	83.30%	9 666 000
1979	5 915 168	6 757 152	87.50%	6 702 760	88.20%	9 857 000
1980	5 917 355	6 925 243	85.40%	6 731 320	87.90%	9 899 000
1983	5 629 996	7 159 349	78.60%	7 271 280	77.40%	10 099 000
1985	5 744 321	7 621 504	75.40%	7 207 920	79.70%	10 011 000
1987	5 623 128	7 741 149	72.60%	7 195 680	78.10%	9 994 000
1991	5 735 431	8 462 347	67.80%	7 301 580	78.60%	9 867 000
1995	5 904 854	8 906 608	66.30%	7 463 960	79.10%	9 821 000

Podem-se identificar três causas principais para o inflacionamento dos cadernos eleitorais em Portugal:

Duplicações de recenseamento – ao mudarem de residência, e consequentemente de freguesia de recenseamento, os eleitores, uma vez na nova freguesia de residência, procedem ao processo de recenseamento. Compete a esta nova freguesia enviar cópia do processo à freguesia de origem, para que aqui seja dada baixa do eleitor. Acontece porém que nem sempre isso se verifica, ou pela morosidade do processo ou mesmo porque a freguesia original pura e simplesmente não abate os seus "emigrados".

Cidadãos falecidos ainda constantes nos cadernos eleitorais – muitos cidadãos já falecidos permanecem nos cadernos eleitorais, quer porque a base de dados do Ministério da justiça onde está concentrado este tipo de informação não está nunca, e compreensivelmente, devidamente actualizada; quer porque não esteve nunca agilizado o processo de transferência de informação entre a base de dados do Ministério da Justiça e as freguesias de recenseamento dos respectivos cidadãos.

Emigrantes – muitos cidadãos portugueses que hoje exercem a sua profissão noutro País estão ainda recenseados em freguesias do Continente e Regiões Autónomas. Talvez esta seja até a causa que melhor explica um tão baixo número de emigrantes recenseados.

Foi tendo presente este conjunto de problemas e a necessidade de conferir maior clareza à base de dados eleitoral, instrumento fulcral do processo democrático, que o Ministro da Administração Interna Alberto Costa propôs à Assembleia da República a Lei que define sob que formas

deveria decorrer o processo de actualização do recenseamento, cuja primeira fase recentemente se concluiu.

Como se evitará um inflacionamento artificial dos cadernos eleitorais, ou seja, como limpar os cadernos eleitorais?
Em primeiro lugar, optou-se por **centralizar** a informação relativa ao recenseamento. Com esta operação, desde logo evitam-se duplicações de inscrições.
Através duma interconexão entre a nova base de dados eleitoral e a base de dados do Ministério da Justiça consegue-se a **eliminação de muitos dos mortos** ainda constantes na base de dados do recenseamento.
A eliminação dos **emigrantes** (não residentes no território nacional, mas recenseados) é uma questão mais complexa e cuja resolução só seria realmente possível e feita de forma eficaz, se os referidos emigrantes se recenseassem nas Embaixadas e Consulados nos países para onde emigraram.

ACTUALIZAÇÃO DOS CADERNOS ELEITORAIS: CONSEQUÊNCIAS PROVÁVEIS

Considerando os dados fornecidos pelas estimativas do Instituto Nacional de Estatística relativos ao ano de 1995, a distribuição dos deputados pelos círculos (calculada pelo método de Hondt, tal como previsto no artigo 13 da lei eleitoral para a Asembleia da República - 14/79) viria a ser a que consta do Quadro III.

QUADRO III

DISTRITO	REC. 95 Nº DEP.	ESTIM. 95 Nº DEP.	DIFERENÇA
AVEIRO	14	15	1
BEJA	4	3	-1
BRAGA	16	17	1
BRAGANÇA	4	3	-1
C. BRANCO	5	5	0
COIMBRA	10	10	0
ÉVORA	4	4	0
FARO	8	8	0
GUARDA	4	4	0
LEIRIA	10	10	0
LISBOA	50	50	0
PORTALEGRE	3	3	0
PORTO	37	38	1
SANTARÉM	10	10	0
SETÚBAL	17	17	0
V. CASTELO	6	5	-1
V. REAL	5	5	0
VISEU	9	9	0
AÇORES	5	5	0
MADEIRA	5	5	0
TOTAL	226	226	0

Verifica-se que, a dar como válidos aqueles números, os distritos de Beja, Bragança e Viana do Castelo teriam perdido nas eleições de Outubro de 1995 um deputado a favor dos de Braga, Porto e Aveiro.

Para além das consequências no número de deputados em cada círculo distrital, haveria que considerar as consequências ao nível da distribuição dos deputados entre os vários

partidos, para cada círculo. Feita a simulação com esta nova distribuição de deputados por círculo, e utilizando os dados (votações) das eleições legislativas de Outubro de 1995, obtivémos a composição da Assembleia da República, tal como descrito no Quadro IV, na coluna "estimativa 95". Como se poderá constatar, no cômputo geral, o Partido Socialista e o Partido Comunista perderiam um deputado, que seriam ganhos pelo Partido Social Democrata e pelo Partido Popular.

QUADRO IV

		1995					ESTIMATIVA 1995			
DISTRITO	Nº DEPUTADOS 95	PS	PSD	PP	PCP	ESTIMATIVA 95	PS	PSD	PP	PCP
AVEIRO	14	6	6	2		15	6	7	2	
BEJA	4	2	1		1	3	2	1		
BRAGA	16	8	7	1		17	8	7		
BRAGANCA	4	2	2			3		2		
C. BRANCO	5	3	2			5	3	2		
COIMBRA	10	6	4			10	6	4		
ÉVORA	4	2	1		1	4	2	1		1
FARO	8	5	3			8	5	3		
GUARDA	4	2	2			4	2	2		
LEIRIA	10	4	5	1		10	4	5	1	
LISBOA	50	24	15	5	6	50	24	15	5	6
PORTALEGRE	3	2	1			3	2	1		
PORTO	37	18	14	3	2	38	19	14	3	2
SANTARÉM	10	5	3	1	1	10	5	3	1	1
SETÚBAL	17	9	3	1	4	17	9	3	1	4
V. CASTELO	6	3	3			5		3		
V. REAL	5	2	3			5	2	3		
VISEU	9	4	4	1		9	4	4	1	
AÇORES	5	2	3			5	2	3		
MADEIRA	5	2	3			5	2	3		
TOTAL	226	111	85	15	15	226				
		226					226			

4. CONCLUSÃO

A actual geografia eleitoral é afectada negativamente pela existência de cadernos eleitorais desactualizados: Estes integram duplas inscrições para além de que contêm indevidamente um número significativo de cidadãos emigrantes e até muitos já falecidos.

Este conjunto de incorrecções origina enviezamentos de vários tipos.

Em primeiro lugar, ao nível da proporcionalidade na distribuição de mandatos por círculos, favorecendo distritos que, sendo menos populosos, comportam nos seus cadernos eleitorais um número considerável de eleitores "fantasma". Com a actualização dos cadernos eleitorais, prevê-se que os distritos de Beja, Bragança e Viana do Castelo percam um deputado cada. Este efeito será tanto mais acentuado, quanto se prevê também a redução do número de deputados na Assembleia da República.

Ao nível dos vários partidos com representação parlamentar, as consequências duma actualização do recenseamento serão reduzidas, uma vez que apenas se prevêem pequenas variações de mais ou menos um deputado por partido, em termos gerais - sendo que estas resultam de variações mais amplas a nível de cada um dos diversos círculos distritais, mas cujos efeitos se contrariam. Podemos pois concluir que na situação presente, e mesmo com cadernos eleitorais desactualizados, se salvaguarda o princípio de proporcionalidade estipulado pelo artigo 288 da Constituição Portuguesa.

ANEXOS

ANEXO 1

$$\text{var}\left(\hat{\text{T}}\right) = \frac{N^2}{n^2} \quad \text{var}\left(\sum_{i=1}^{N} X_i \delta_i\right)$$

$$= \frac{N^2}{n^2}\left[\sum_{i=1}^{N} X_i^2 \quad \text{var}\left(\delta_i\right) + 2\sum_{i=1}^{N}\sum_{j>1}^{N} X_i X_j \quad \text{cov}\left(\delta_i, \delta_j\right)\right]$$

Mas $\quad \text{var}\left(\delta_i\right) = \frac{n}{N}\left(1 - \frac{n}{N}\right) = \frac{n}{N}\left(\frac{N-n}{N}\right)$

e

$$\text{cov}\left(\delta_i, \delta_j\right) = E\left(\delta_i, \delta_j\right) - \left(\frac{n}{N}\right)^2$$

$$= \frac{n}{N}\frac{n-1}{N-1} - \left(\frac{n}{N}\right)^2$$

$$= \frac{n}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right) = \frac{n}{N}\left(\frac{n-N}{N(N-1)}\right)$$

$$= -\frac{n(N-n)}{N^2(N-1)}$$

$$\Rightarrow \text{var}\left(\hat{\text{T}}\right) = \frac{N^2}{n^2}\left[\frac{n}{N}\left(\frac{N-n}{N}\right)\left(\sum X_i^2\right) - \frac{n}{N}\left(\frac{N-n}{N(N-1)}\right)\cdot\left(2\sum\sum X_i X_j\right)\right]$$

$$= \frac{N^2}{n^2}\frac{n}{N}\left(1-f\right)\left[\left(\sum X_i^2\right) - \frac{2}{N-1}\sum\sum X_i X_j\right]$$

$$= \frac{N^2}{n^2}\frac{n}{N}\left(1-f\right)\frac{N\sum_{i=1}^{N}\left(X_i - u\right)^2}{N-1} = N^2\left(1-f\right)\frac{\sigma'^2}{n}$$

ANEXO 2

$$\underset{2|1}{\mathrm{V}}\left(\hat{\mathrm{T}}\right) = \underset{2|1}{\mathrm{V}}\left[M \sum_{i\in s} \frac{\hat{\mathrm{T}}_i}{m} \right]$$

$$= \left(\frac{M}{m}\right)^2 \sum_{i\in s} \underset{2|1}{\mathrm{V}}\left(\hat{\mathrm{T}}_i\right)$$

$$= \left(\frac{M}{m}\right)^2 \sum_{i\in s} N_i^2 \left(1-f_i\right) \frac{\sigma'^2_{2,i}}{n_i}$$

com $\quad \sigma'^2_{2,i} = \dfrac{1}{N_i - 1} \sum_{j=1}^{N_i} \left(x_{ij} - \mu_i\right)^2$

logo $\quad \underset{1}{\mathrm{E}}\left(\underset{2|1}{\mathrm{V}}\left(\hat{\mathrm{T}}\right)\right) = \dfrac{M^2}{m^2} \underset{1}{\mathrm{E}}\left(\sum_{i\in s} N_i^2 \left(1-f_i\right) \dfrac{\sigma'^2_{2,i}}{n_i} \right)$

Mas $\quad \sum_{i\in s} N_i^2 \left(1-f_i\right) \dfrac{\sigma'^2_{2,i}}{n_i} = \sum_{i=1}^{M} N_i^2 \left(1-f_i\right) \dfrac{\sigma'^2_{2,i}}{n_i} \delta_i$

$$\text{com } \delta_i = \begin{cases} 1 & \text{se } i \in s \\ 0 & \text{se } i \notin s \end{cases}$$

Logo

$$\mathrm{E}\left(\delta_i\right) = P\left(\delta_i = 1\right) = \frac{m}{M}$$

$$\Rightarrow \underset{1}{\mathrm{E}}\left(\underset{2|1}{\mathrm{V}}\left(\hat{\mathrm{T}}\right)\right) = \frac{M}{m} \sum_{i=1}^{M} N_i^2 \left(1-f_i\right) \frac{\sigma'^2_{2,i}}{n_i}$$

Por outro lado

$$\begin{aligned}
\underset{2|1}{\mathrm{E}}\left(\hat{\mathrm{T}}\right) &= \underset{2|1}{\mathrm{E}}\left(\frac{M}{m}\sum_{i\in s}\underset{2|1}{\mathrm{E}}\left(\hat{\mathrm{T}}_i\right)\right) \\
&= \frac{M}{m}\sum_{i\in s}\mathrm{T}_i \\
\Rightarrow \underset{1}{\mathrm{V}}\left[\underset{2|1}{\mathrm{E}}\left(\hat{\mathrm{T}}\right)\right] &= M^2\,\underset{1}{\mathrm{V}}\left(\frac{\sum_{i\in s}\mathrm{T}_i}{m}\right) \\
&= M^2\left(1-\frac{m}{M}\right)\left(\frac{\sigma_1^2}{m}\right)
\end{aligned}$$

ANEXO 3

Sendo $\hat{T} = \sum_{i=1}^{N} \frac{X_i}{p_i} \delta_i$

onde

$$E(\delta_i) = p_i$$
$$v(\delta_i) = p_i(1 - p_i)$$

e

$$\operatorname{cov}\left(\delta_i, \delta_i\right)_{i \neq j} = E(\delta_i, \delta_j) - E(\delta_i)E(\delta_j)$$
$$= p_{ij} - P_i P_j$$

$$v(\hat{T}) = \sum_{i=1}^{N} \left(\frac{X_i}{p_i}\right)^2 v(\delta_i) + \sum_{i=1}^{N} \sum_{j \neq 1}^{N} \frac{X_i}{p_i} \operatorname{cov}(\delta_i, \delta_j)$$
$$= \sum_{i=1}^{N} \frac{X_i^2}{p_i}(1 - p_i) + \sum_{i=1}^{N} \sum_{j \neq 1}^{N} \frac{X_i}{p_i} \frac{X_j}{p_j} \left(P_{ij} - P_i P_j\right) .$$

ANEXO 4

1. Introdução

Em 1998, tem início uma nova série do Inquérito ao Emprego (IE), cujas alterações relativamente à série anterior visam sobretudo a adaptação às alterações aprovadas pelo *EUROSTAT* para o Inquérito Comunitário.

De acordo com as directivas comunitárias, tendo por objectivo a harmonização dos inquéritos realizados nos diversos países, o IE apresenta alterações significativas no que respeita ao plano de amostragem, dimensão e rotação da amostra e a nível do questionário.

A amostra é constituída por secções de igual dimensão em cada região NUTS II, distribuídas uniformemente pelas treze semanas do trimestre, de forma a permitir apuramentos de resultados intercalares. A extrapolação dos dados é feita através de ponderadores por sexo e idade, para além da ponderação por região NUTS II já utilizada na anterior série.

O questionário tem uma estrutura diferente da anterior, tendo o conteúdo sofrido igualmente algumas alterações, embora globalmente incida sobre os mesmos temas e as questões se baseiem em conceitos semelhantes aos da anterior série. A hipótese inicialmente prevista pelo *EUROSTAT* de estruturar o questionário de forma modular, de forma a inquirir alguns módulos apenas num dos trimestres do ano, foi abandonada por se considerar que daí não advinham benefícios em termos de qualidade do inquérito nem redução de custos significativos.

Mantém-se a total observância das recomendações internacionais, sendo que a **principal alteração conceptual é a passagem da idade mínima da população activa dos 14 para os 15 anos. Tal alteração deriva do facto da idade mínima estar intimamente associada ao término da escolaridade obrigatória.** De qualquer forma, e à semelhança da anterior série, está prevista a inquirição regular do módulo relativo à actividade económica aos indivíduos com idade inferior a 15 anos, tendo por objectivo avaliar a dimensão do fenómeno do "**trabalho infantil**".

2 - Desenho da amostra

2.1 - Unidade de amostragem

Considerou-se como unidade de amostragem o alojamento, por ser uma unidade estatística estável, de fácil identificação e relativamente simples de actualizar.

2.2 - Base de amostragem

A base de amostragem é a Amostra-Mãe. A Amostra-Mãe é uma amostra de unidades de alojamento destinada a ser utilizada pelos inquéritos a realizar junto das famílias pelo Instituto Nacional de Estatística, entre os quais se inclui o Inquérito ao Emprego.

3 - Amostra-Mãe

3.1 - Desenho da Amostra-Mãe

Atendendo à unidade de amostragem considerada - o alojamento - seguiu-se o método de amostragem probabilística e multietápica habitualmente usado neste tipo de amostras.

Tomaram-se como unidades primárias as freguesias, pois constituem as unidades administrativas mais pequenas, com limites físicos bem definidos, de que se disporia de informação sobre o número de unidades de alojamento a fornecer pelos Censos 91.

Tendo em conta a informação da Base Geográfica de Referenciação Espacial, seleccionaram-se numa 2ª etapa, nas freguesias da amostra - secções estatísticas -, áreas geográficas contíguas, delimitadas cartograficamente, com cerca de 300 alojamentos.

Sendo a Amostra-Mãe constituída por todos os alojamentos incluídos nas secções da amostra, foi efectuada a operação de listagem das suas unidades de alojamento, cada uma delas referenciada pela respectiva morada e nome do representante da família.

Posteriormente, para cada inquérito utilizador, é construída uma subamostra que pressupõe uma 3ª etapa em que são seleccionados os

alojamentos a inquirir, de modo a perfazer a dimensão global da amostra previamente definida.

Em 1992, a dimensão da Amostra-Mãe foi calculada de modo a satisfazer as directivas do *EUROSTAT* então em vigor para o IE, acrescida de uma estimativa dos alojamentos necessários para outros inquéritos que se previa efectuar junto das famílias até ao Censo seguinte.

A representatividade da Amostra-Mãe é o nível NUTS II, tendo sido seleccionada independentemente em cada uma das sete regiões: Norte, Centro, Lisboa e Vale do Tejo, Alentejo, Algarve, Região Autónoma dos Açores e Região Autónoma da Madeira.

3.1.1 - Selecção das freguesias (unidades primárias)

Na primeira etapa, foram seleccionadas as freguesias de modo sistemático e com probabilidade proporcional ao número de eleitores.

A informação mais actualizada de que se dispunha à data sobre a dimensão das freguesias era a do Recenseamento Eleitoral, porque os dados dos Censos 91 ainda não tinham sido apurados.

Calculada a dimensão da Amostra-Mãe em termos de número de alojamentos, efectuou-se a sua repartição proporcional pelas sete regiões NUTS II, a partir das estimativas da população e do número médio de indivíduos por alojamento.

O número de secções a seleccionar por região foi determinado em função do número de alojamentos necessários para a Amostra-Mãe, entrando em linha de conta com a dimensão média das secções estatísticas (cerca de 300 unidades de alojamento).

Em cada uma das regiões, os concelhos foram ordenados em serpentina, tendo-se procedido posteriormente à ordenação alfabética das freguesias dentro do concelho. Obteve-se uma lista, por região, na qual se seleccionaram sistematicamente as freguesias, depois de determinar o respectivo intervalo de selecção.

A probabilidade de selecção da freguesia i da região r é:

$$P_{ri} = \begin{cases} \dfrac{E_{ri}}{I_r} & se \quad E_{ri} < I_r \\ \\ 1 & se \quad E_{ri} \geq I_r \end{cases} \qquad r = 1,2,3,4,5,6,7 \qquad (4.1)$$

em que,

E_{ri} representa o número de residentes da freguesia i da região r,

I_r representa o intervalo de selecção sistemático da região r,

$I_r = \dfrac{\sum_i E_{ri}}{S_r}$ em que S_r representa o número de secções a seleccionar na região r.

Nas freguesias com dimensão superior ao intervalo de selecção calculado, o número de secções a seleccionar depende da respectiva dimensão e fica determinado pela selecção sistemática. Nas restantes freguesias, é seleccionada uma única secção.

3.1.2 - Selecção das secções (unidades secundárias)

Na segunda etapa, foram seleccionadas sistematicamente as secções usando a informação da Base Geográfica de Referenciação Espacial, uma vez que não estava ainda concluído o trabalho de campo dos Censos 91.

Nas freguesias em que seria seleccionada mais do que uma secção, houve necessidade de calcular o intervalo de selecção sistemático, a utilizar na 2ª etapa, a partir da taxa média t_m de eleitores por alojamento, ou seja,

$$I_{rj} = \frac{I_r}{t_m}$$ em que I_r é o intervalo de selecção sistemático utilizado para a selecção das freguesias da região r.

Em função da dimensão da freguesia, a probabilidade de selecção da secção j, da freguesia i, da região r, é dada por:

$$P_{rij} = \begin{cases} \dfrac{A_{ri}}{I_r} & se \quad E_{ri} < I_r \\ & r = 1,2,3,4,5,6,7 \\ \dfrac{A_{rij}}{\sum_j A_{rij}} & se \quad E_{ri} \geq I_r \end{cases} \qquad (4.2)$$

em que,

A_{rij} é o número de unidades de alojamento da secção j, da freguesia i, da região r, segundo a Base Geográfica de Referenciação Espacial,

$\sum_j A_{rij}$ representa o número total de unidades de alojamento da secção j, da freguesia i, da região r, fornecido pela fonte acima referida.

O número total de secções da amostra é 1143, cuja distribuição pelas regiões NUTS II figura num dos parágrafos seguintes, onde consta a dimensão da amostra do IE por região e trimestre.

Em cada trimestre todas as secções estão representadas na amostra.

3.1.3 - Selecção de subamostras

Em geral, para cada inquérito a efectuar junto das famílias, é seleccionada uma subamostra autoponderada da Amostra-Mãe, ou seja, em que cada alojamento tem igual probabilidade de selecção na respectiva região NUTS II.

A selecção dos alojamentos é efectuada numa 3ª etapa, constituindo estes as unidades terciárias.

A opção pela autoponderação tem como principais vantagens:

a) possibilitar o tratamento dos dados obtidos da amostra como provenientes de um censo, o que facilita o cálculo dos estimadores e suas variâncias;

b) simplificar a planificação e supervisão do trabalho de campo, pois será inquirido aproximadamente o mesmo número de unidades de alojamento em cada uma das secções seleccionadas.

Para uma subamostra ser autoponderada, a probabilidade final de selecção de cada alojamento, que é dada pelo produto das probabilidades de selecção em cada uma das três etapas, terá de ser constante. Esta taxa de amostragem é determinada em função do número de alojamentos que se pretendem para cada inquérito, a fim de satisfazer as restrições de precisão e/ou de custo previamente definidas.

A probabilidade final de selecção dos alojamentos, em cada uma das regiões NUTS II, é dada por:

$$P_{rijk} = P_{ri}\ P_{rij}\ P_{rijk} = t_r \qquad \textbf{(4.3)}$$

sendo:

t_r a taxa de amostragem determinada para a região r, dada por:

$t_r = \dfrac{n_{rij}}{N_{rij}}$ em que n_{rij} representa o número de unidades de alojamento a seleccionar na secção j, da freguesia i, na região r,

N_{rij} representa o número de unidades de alojamento listadas da secção j, da freguesia i, na região r.

O número de alojamentos a seleccionar por secção, em cada região r, é obtido mediante a resolução da equação acima, ou seja,

$$n_{rij} = \frac{t_r \times N_{rij}}{P_{ri} \times P_{rij}}. \qquad (4.4)$$

Por uma questão de redução de custos, a selecção dos alojamentos é feita em bloco, isto é, escolhido um alojamento da secção, são seleccionados os que lhe seguem pela ordem em que figuram na listagem da secção.

3.2 Actualização da Amostra-Mãe

A Amostra-Mãe, construída a partir dos dados dos Censos de 91, foi inicialmente dimensionada de modo a dar resposta no período intercensitário às necessidades do Inquérito ao Emprego, assim como a outros inquéritos, nomeadamente o Inquérito aos Orçamentos Familiares.

Pelo esgotamento parcial da Amostra-Mãe, pela sua desactualização natural em cinco anos, derivada quer das alterações detectadas, quer pela não inclusão das novas construções e tomando em consideração a remodelação do IE em 1998, foi decidido proceder a uma actualização no campo.

A operação de actualização da Amostra-Mãe, que decorreu no ano de 1996, consistiu na substituição das 1143 secções, seleccionadas em 1992, por secções contíguas, pertencentes à mesma freguesia ou à freguesia geograficamente mais próxima, nos casos em que a freguesia era constituída por uma única secção.

4 - Estimadores

4.1 - Estimador do total

O estimador do total relativo a uma determinada característica, para uma dada região geográfica, é obtido através da expressão genérica:

$$\hat{T}_X = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i} T_y^*, \qquad (5.1)$$

em que,

$\sum_{i=1}^{n} X_i$ representa o total de uma característica na região geográfica, obtida a partir dos valores da amostra,

$\sum_{i=1}^{n} Y_i$ representa o total de indivíduos da amostra, na região geográfica,

T_y^* **representa a estimativa independente da população na região em causa, relativamente ao ponto médio do período de referência do inquérito,**

$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i}$ representa a proporção de indivíduos na amostra da região geográfica, com uma determinada característica.

A estimativa de um total para uma agregação de regiões obtém-se adicionando as respectivas estimativas.

Encontra-se em fase de implementação uma estratificação *à posteriori*, para correcção de eventuais deficiências da amostra, relativas à distribuição dos indivíduos por sexo e escalão etário, uma vez que a unidade de amostragem é o alojamento.

As estimativas serão afectadas por um ponderador que fará a correcção das discrepâncias detectadas na amostra relativamente à população.

4.2 - Precisão dos estimadores - Erros de amostragem

A precisão do estimador do total $\hat{T}_X$, ou seja, o seu erro de amostragem, é avaliado em termos relativos pelo coeficiente de variação, expresso em percentagem e obtido através da expressão:

$$\widehat{c.v.}\left(\hat{T}_X\right) = \frac{\sqrt{\widehat{\text{var}}\left(\hat{T}_X\right)}}{\hat{T}_X} \times 100\,\%\,, \qquad (5.2)$$

sendo,

$$\widehat{\text{var}}\left(\hat{T}_X\right) = \left(T_y^*\right)^2 \widehat{\text{var}}\left(\hat{p}\right) \qquad (5.3)$$

$$\widehat{\text{var}}(\hat{p}) = \sum_{i=1}^{N} y_i^2 \left[\widehat{\text{var}}\left(\sum_{i=1}^{n} X_i\right) + \left(\hat{p}\right)^2 \widehat{\text{var}}\left(\sum_{i=1}^{n} Y_i\right) - 2\hat{p}\,\widehat{\text{cov}}\left(\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i\right) \right] \qquad (5.4)$$

em que $\widehat{\text{var}}\left(\sum_{i=1}^{n} X_i\right)$, $\widehat{\text{var}}\left(\sum_{i=1}^{n} Y_i\right)$ e $\widehat{\text{cov}}\left(\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i\right)$ foram calculadas pelo método das diferenças sucessivas, ou seja,

$$\widehat{\text{var}}\left(\sum_{i=1}^{n} X_i\right) = \frac{n}{2(n-1)} \sum_{i=1}^{n-1} \left(x_i - x_{i+1}\right)^2\,, \qquad (5.5)$$

$$\widehat{\mathrm{var}}\left(\sum_{i=1}^{n} Y_i\right) = \frac{n}{2(n-1)} \sum_{i=1}^{n-1} \left(y_i - y_{i+1}\right)^2 , \qquad (5.6)$$

$$\widehat{\mathrm{cov}}\left(\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i\right) = \frac{n}{2(n-1)} \sum_{i=1}^{n-1} \left(x_i - x_{i+1}\right)\left(y_i - y_{i+1}\right), \qquad (5.7)$$

em que,

n é o número de secções da amostra de uma determinada desagregação geográfica,

5 - Amostra do Inquérito ao Emprego

5.1 - Dimensão da amostra

De acordo com as directivas comunitárias, a amostra para a nova série do IE apresenta algumas alterações em relação à da série anterior, no que respeita ao plano de amostragem, dimensão e rotação da amostra.

A dimensão da amostra do novo IE foi determinada de modo a satisfazer os requisitos de precisão definidos pelo *EUROSTAT.*

As condições de precisão referem-se às estimativas anuais e às estimativas das variações entre dois trimestres consecutivos, para subpopulações com pelo menos 5% da população em idade activa. Foi então estipulado que a dimensão da amostra, o plano de amostragem e a rotação da amostra devem satisfazer:

1. Precisão em nível - o desvio padrão das estimativas anuais, ao nível das regiões NUTS II com pelo menos 300 000 habitantes, não deve exceder 8% das subpopulações em causa.

2. Precisão em evolução - o desvio padrão das estimativas das variações entre trimestres consecutivos, ao nível nacional , não deve exceder 3% das mesmas subpopulações.

Em cada região, o número de alojamentos a inquirir por trimestre em cada secção será idêntico, de modo a distribuir uniformemente o trabalho de campo pelos três meses que constituem o trimestre.

Para o estudo do dimensionamento da amostra, utilizaram-se as estimativas da precisão do estimador do quociente associado ao estimador do total. O cálculo destas estimativas foi efectuado com o software *QUANTUM*, utilizando para o efeito os resultados do 3° e 4° trimestres de 1996 do IE.

As fórmulas para o cálculo dos desvios-padrão baseiam-se no método das diferenças sucessivas, processo que se utiliza habitualmente na prática para calcular erros de médias e de proporções, obtidos a partir de amostras estratificadas multietápicas.

5.1.1 - Dimensionamento para a precisão em nível

Face aos valores dos coeficientes de variação calculados a partir dos resultados do 3° trimestre de 1996 do IE, dimensionou-se a amostra, para cada uma das regiões NUTS II, de modo a obter a precisão em nível previamente definida.

Assim, para as regiões do Norte, Centro e Lisboa e Vale do Tejo, determinou-se a dimensão da amostra necessária para que o coeficiente de variação das estimativas trimestrais de subpopulações com pelo menos 5% de população activa, não ultrapasse os 8%. As regiões do Alentejo, Algarve e as Regiões Autónomas dos Açores e da Madeira, por terem menos de 300 000 habitantes, estão isentas dessa obrigação; no entanto, a dimensão atribuída garante, para as estimativas referentes às mesmas subpopulações, coeficientes de variação não significativamente superiores a 10%.

Para calcular a dimensão da amostra de modo a garantir a precisão pretendida, a metodologia utilizada foi a seguinte:

1. Pressupondo a utilização de amostragem aleatória simples, estimou-se para determinadas variáveis o número de indivíduos a entrevistar,

$$n_a = \frac{q}{p\,(cv)^2}, \qquad (6.1)$$

sendo,

n_a dimensão da amostra em número de indivíduos, para amostragem aleatória simples, necessária para garantir uma dada precisão (coeficiente de variação *cv*),

p proporção de indivíduos na amostra do IE, do 3º trimestre de 1996, com a característica em causa,

$q = 1 - p$

cv coeficiente de variação que se pretende admitir para a proporção, em função da região em causa.

2. Para as mesmas variáveis, calculou-se o valor do *Deff*, efeito do desenho da amostra, dado pelo quociente entre as variâncias do respectivo estimador, utilizando amostragem aleatória simples e amostragem por conglomerados,

$$Deff = \frac{\mathrm{var}_c\,(p)}{\mathrm{var}_a\,(p)}, \qquad (6.2)$$

em que,

$\mathrm{var}_c\,(p)$ variância do estimador da proporção utilizando amostragem por conglomerados, calculada através do software *QUANTUM*, fórmula (5.4) ,

$\mathrm{var}_a\,(p)$ variância do estimador da proporção utilizando amostragem aleatória simples,

$$\mathrm{var}_a(p) = \frac{p\,q}{n}$$

n dimensão da amostra (número de indivíduos entrevistados).

3. Atendendo a que o *Deff* é o factor pelo qual é necessário multiplicar a dimensão de uma amostra aleatória simples, para obter a mesma precisão com uma amostra por conglomerados, com o mesmo número de elementos, determinou-se para o tipo de amostragem adoptado no IE o número de indivíduos a inquirir, mediante a expressão:

$$n_c = n_a \times Deff\ , \qquad (6.3)$$

em que,

n_c dimensão da amostra de indivíduos utilizando amostragem por conglomerados,

n_a dimensão da amostra de indivíduos utilizando amostragem aleatória simples.

4. Assumindo a relação entre alojamentos clássicos e população residente encontrada na amostra do trimestre em estudo, em cada uma das regiões, determinou-se o número de unidades de alojamento necessário para garantir, em cada uma das situações, coeficientes de variação não superiores aos pretendidos:

$$n_c^{UA} = \frac{Alojamentos\ Classicos\ da\ região\ R}{População\ Residente\ na\ região\ R}\ n_c^{IE96}\ , \qquad (6.4)$$

em que,

n_c^{UA} dimensão da amostra de unidades de alojamento utilizando amostragem por conglomerados,

n_c dimensão da amostra de indivíduos calculada para os coeficientes de variação considerados, utilizando amostragem por conglomerados,

n_c^{IE96} número de indivíduos inquiridos no 3° trimestre de 1996 do Inquérito ao Emprego.

5.1.2 - Dimensionamento para a precisão em evolução

Para o cálculo da precisão em evolução, assumiu-se a utilização de amostragem aleatória simples, ignorando-se quaisquer efeitos devidos ao desenho da amostra por constituição de conglomerados ou estratos.

A precisão da evolução das estimativas a nível nacional relativamente à população activa (erro relativo - *er*), em trimestres consecutivos, foi determinada mediante a seguinte fórmula:

$$er(|p2 - p1|) = \sqrt{\frac{2p_1q_1}{n}(1 - \rho\ p_{12})} \Big/ p_1 \times 100\ \%\ , \tag{6.5}$$

em que,

$\rho = \dfrac{\mathrm{cov}(p_1, p_2)}{\sqrt{\mathrm{var}(p_1)\,\mathrm{var}(p_2)}}$ é o coeficiente de correlação linear da proporção da característica nos dois trimestres em causa,

p_1 e p_2 são respectivamente as proporções estimadas, a nível nacional, de indivíduos em idade activa com uma determinada característica, para os dois trimestres em estudo,

p_{12} é a proporção de indivíduos em idade activa, comuns aos dois trimestres.

A resolução da equação (6.5), permitiu confirmar que a dimensão da amostra definida para a precisão em evolução garantia, para as variáveis com peso não inferior a 5% da população em idade activa, erros relativos não superiores a 3%.

Por região e por trimestre, o número de secções, a dimensão da amostra e o número de alojamentos a inquirir por secção são os seguintes:

Região	**Nº de Aloj./Secção**	**Nº de Secções**	**Nº de Alojamentos**
Norte	17	326	5 542
Centro	14	183	2 562
Lisboa V. Tejo	15	345	5 175
Alentejo	29	72	2 088
Algarve	32	75	2 400
R.A. Açores	20	72	1 440
R.A. Madeira	22	70	1 540
		1 143	20 747

Como as directivas do *EUROSTAT* exigiam que o inquérito deveria ser contínuo, a amostra foi distribuída pelas 13 semanas de trabalho de cada trimestre, a fim de permitir apurar resultados trimestrais. Será previamente definida a semana de referência a que cada entrevista se deve reportar, nas diversas questões presentes no questionário, devendo a mesma ser efectuada, sempre que possível, na semana seguinte.

5.2 - Rotação da amostra

Por compromisso entre a redução do enviesamento, provocado pela manutenção dos mesmos alojamentos na amostra durante um período prolongado e o rápido esgotamento da amostra, em consequência da renovação de todos os alojamentos dum trimestre para o outro, optou-se por manter o esquema de rotação de 1/6 da amostra em cada trimestre.

Houve necessidade de contemplar no esquema de rotação o facto de o inquérito ter passado a ser contínuo. Em cada uma das secções, 1/6 dos alojamentos será substituído pelo mesmo número de alojamentos da secção, garantindo a cobertura regional da amostra em quaisquer treze semanas consecutivas.

ANEXO 5

Lei nº 31/91 de 20 de Julho

Publicação e difusão de sondagens e inquéritos de opinião

A Assembleia da República decreta, nos termos dos artigos 164º., alínea *d*), e 169., nº. 3, da Constituição, o seguinte:

Artigo 1º.
Objecto

1 - A presente lei regula a realização de sondagens e inquéritos de opinião destinados a publicação ou difusão em orgãos de comunicação social, cujo objecto se relacione directa ou indirectamente com a realização de qualquer acto eleitoral para os orgãos de soberania, das Regiões Autónomas, das autarquias locais e para o Parlamento Europeu ou de referendo nacional ou local, bem como a sua publicação ou difusão.
2 - É abrangida pelo disposto no número anterior a publicação ou difusão de previsões ou simulações de voto que se baseiem em sondagens ou inquéritos relativos a qualquer acto eleitoral ou referendário nele referidos.

Artigo 2º.
Inscrição

As sondagens e inquéritos de opinião referidos no nº. 1 do artigo anterior só podem ser realizados por entidades que se tenham inscrito para o exercício desta actividade junto da Alta Autoridade para a Comunicação Social.

Artigo 3º.
Regras a observar

A realização de sondagens e inquéritos de opinião deve obedecer às seguintes regras:

a) A amostragem deve ser representativa do universo a abranger;
b) As perguntas devem ser formuladas com objectividade, clareza e precisão e sem sugerirem, directa ou indirectamente, o sentido das respostas;
c) A duração do inquérito deve ser curta e permitir a homogeneidade dos resultados;
d) Os inquiridos devem ser informados de qual a entidade responsável pela sondagem;
e) **Deve ser preservada a identidade das pessoas inquiridas, bem como as suas respostas;**
f) A interpretação dos resultados brutos deve ser feita de forma a não falsear ou deturpar o resultado da sondagem.

Artigo 4º.
Depósito

A entidade responsável pela publicação ou difusão da sondagem ou inquérito deve proceder ao seu depósito junto da Alta Autoridade para a Comunicação Social até ao dia da publicação ou difusão.

Artigo 5º.
Ficha técnica

O depósito de sondagem a que se refere o artigo anterior é acompanhado de uma ficha técnica onde conste obrigatoriamente:

a) Identificação da entidade que realizou a sondagem;
b) Identificação do cliente;
c) Objecto da sondagem ou inquérito;
d) Descrição do universo abrangido e sua quantificação;
e) Número de pessoas inquiridas (amostra), sua repartição geográfica e composição, evidenciando-se a **amostra prevista e a obtida**;
f) Descrição da metodologia de selecção da amostra, referenciando-se os métodos sucessivos de selecção de unidades até aos inquiridos;
g) Indicação da técnica de recolha de informação (postal, telefónica, pessoal ou outra);
h) Indicação dos métodos de controlo da recolha de informação e percentagem de entrevistas controladas;
***i)* Taxa de resposta e indicação de eventuais enviesamentos que os não respondentes possam introduzir;**
j) Texto integral das questões colocadas;
k) Margem de erro máximo associado a cada ventilação;
l) Métodos e coeficientes máximos de ponderação eventualmente utilizados;
m) Data ou datas em que tiveram lugar os trabalhos de recolha de informação.

Artigo 6º.
Requisitos para a publicação ou difusão

1 - **A primeira publicação ou difusão de sondagens e inquéritos é sempre** acompanhada da publicação da ficha técnica referida no artigo anterior.
2 - Nos restantes casos é obrigatória a publicação dos dados da ficha técnica a que se referem as alíneas a), b), c), d), e), f), k), e m) do artigo anterior.

Artigo 7º.
Primeira publicação ou difusão

A primeira publicação ou difusão de qualquer sondagem ou inquérito de opinião através de orgão de comunicação social deve fazer-se até 10 dias a contar da data do recebimento do depósito obrigatório junto da entidade fiscalizadora.

Artigo 8º.
Publicação ou difusão em períodos eleitorais

Nos sete dias que antecedem o dia da eleição ou de votação para referendo, e até ao encerramento das urnas, são proibidos a publicação, difusão, comentário ou análise de qualquer sondagem ou inquérito de opinião directa ou indirectamente relacionados com o acto eleitoral ou referendário.

Artigo 9º.
Autoridade fiscalizadora

1 – A entidade competente para verificar as condições de realização das sondagens e inquéritos de opinião e o rigor de objectividade na publicação dos resultados, nos termos

definidos pela presente lei, é a Alta Autoridade para a Comunicação Social.
2 - À Comissão Nacional de Eleições cabe a fiscalização ao cumprimento dos disposto no artigo anterior.

Artigo 10º.
Regra de concorrência

A Alta Autoridade para a Comunicação Social deve assegurar que as entidades que realizem sondagens não procedam por acções concertadas, convenções ou entendimentos expressos ou tácitos que tenham como intenção ou procurem ter como efeito impedir ou restringir a mesma actividade a outras entidades.

Artigo 11º.
Dever de colaboração

A entidade que realizou a sondagem ou inquérito deve colocar à disposição da Alta Autoridade para a Comunicação Social todos os documentos e processos na base dos quais a sondagem foi publicada ou difundida, quando isso lhe seja solicitado.

Artigo 12º.
Poderes de verificação

A Alta Autoridade para a Comunicação Social dispõe da faculdade de verificar se as sondagens e inquéritos foram realizados em conformidade com a lei e os regulamentos aplicáveis.

Artigo 13º.
Obrigatoriedade de rectificação

Os orgãos de informação que publicarem ou difundirem qualquer sondagem com violação das disposições da presente lei ou alterando o significado dos resultados obtidos são obrigados a publicar com a mesma relevância as correcções exigidas pela Alta Autoridade para a Comunicação Social e devem dar cumprimento a essa obrigação na publicação seguinte após a notificação dessas correcções.

Artigo 14º.
Contra-ordenações

1 - É punido com coima de montante mínimo de 1 000 000$ e máximo de 5 000 000$, sendo o infractor pessoa singular, ou de 5 000 000$ a 10 000 000$, tratando-se de pessoa colectiva:

a) Quem publicar sondagem ou inquérito em orgão de comunicação social sem que tenha feito o depósito nos termos dos artigos 4º. E 5º.;
b) Quem publicar ou difundir sondagens ou inquéritos sem os requisitos previstos no artigo 6º.;
c) Quem publicar ou difundir sondagens ou inquéritos, bem como o seu comentário ou análise, nos sete dias que antecedem o dia das eleições para orgãos de soberania, Regiões Autónomas e autarquias locais ou para o Parlamento Europeu ou da votação para referendo nacional ou local que directa ou indirectamente se relacione com o acto eleitoral ou votação referidos;
d) Quem, tendo realizado, publicado ou difundido sondagem ou inquérito, não faculte à Alta Autoridade para a Comunicação Social os documentos ou processos por ela solicitados no exercício das suas funções;

e) Quem não der cumprimento ao dever de rectificação previsto no artigo 13°. ou de publicação ou difusão das decisões administrativas ou judiciais a que se refere o artigo 15°.;
f) Quem não der cumprimento ao disposto na alínea *e)* do artigo 3°.

2 - Os limites máximo e mínimo das sanções previstas nos números anteriores podem ser reduzidos a metade, no caso de publicação ou difusão em órgão de comunicação social de sondagens ou inquéritos não acompanhada dos elementos referidos nas alíneas h), I), j), e l), da presente lei.

3 - O produto das coimas reverte integralmente para os cofres do Estado.

Artigo 15°.
Publicação ou difusão das decisões administrativas ou judiciais

A decisão irrecorrida que aplique coima prevista nas alíneas a), b), c), d), e e) do n°. 1 do artigo anterior, ou a decisão judicial transitada em julgado relativa a recurso da mesma decisão, é obrigatoriamente publicada ou difundida pelo destinatário da coima nos termos previstos no artigo 13°.

Artigo 16°.
Norma revogatória

Ficam revogados os artigos 50°. Do Decreto-Lei n°. 319-A/76, de 3 de Maio, 51°. Do Decreto-Lei n°. 701-B/76, de 29 de Setembro, e 60°. Da Lei n°. 14/79, de 16 de maio.

ANEXO 6

SISTEMAS ELEITORAIS
CÍRCULOS UNINOMINAIS
ESCRUTÍNIO UNINOMINAL
SISTEMAS MAIORITÁRIO
UMA VOLTA
DUAS OU MAIS VOLTAS
MAIORIA RELATIVA
MAIORIA ABSOLUTA
SISTEMA MAIORITÁRIO
VOTO ÚNICO
VÁRIOS VOTOS
VOTO ALTERNATIVO
SISTEMA JAPONÊS
VOTO LIMITADO
VOTO BLOQUEADO
VOTO CUMULATIVO
SISTEMAS PREDOMINANTEMENTE MAIORITÁRIO
ESCRUTÍNIO UNINOMINAL
LISTA BLOQUEADA
LISTA COMPLETA
LISTAS COMPOSTAS
SISTEMAS PREDOMINANTEMENTE PROPORCIONAIS
CÍRCULOS PLURINOMINAIS
ESCRUTÍNIO DE LISTA
SISTEMAS PROPORCIONAL
APARENTAMENTOS VOTO PREFERENCIAL
VÁRIOS CÍRCULOS
UM SÓ CÍRCULO
VOTO ÚNICO TRANSFERÍVEL
SIST. DE BARRA
SIST. DINAMARQUÊS
APARENTAMENTO REFORÇADO
NÚMERO UNIFORME
DIVISOR OU MÉDIA MAIS ALTA
HONDT, etc.
QUOCIENTE ELEITORAL
SISTEMAS DE PROPORCIONALIDADE APROXIMADA
SISTEMA PROPORCIONAL INTEGRAL

BIBLIOGRAFIA

- ANTOINE J. (1990), Sondagem, Instrumento de marketing e publicidade, Ed. Cetop
- ARDILLY P. (1994), Les techniques de sondage, Ed. Technif
- BARNET V. (1991), Sample Survey, Principles and methods, Hodder/Stonghton
- BIEMER, PAUL P., GROVES, ROBERT M., MATHIOWETZ, LARSE., SUDMAN, SEYMOUR (1991) Ed. Wiley
- CALDEIRA, OSVALDO (1995), Potencialidades da Amostragem em duas Fases, Ed. ISEGI - Univ. Nova de Lisboa
- COCHRAN W. (1977), Sampling Techniques, 3ª ed., Wiley
- COELHO, PEDRO (1996). Estimadores combinados para pequenos domínios. *Revista de Estatística*, 2, 23-43.
- COTTERET, Jean Marie et ÉMERI, Claude - *Systèmes électoraux.* Paris, Presses Universitaires de France
- CRESPO T. (1993), Técnicas de Amostragens, Ed. CESD
- DROESBEKE J. THOVERON G. (ed.) (1987), Au Rouyaume des Sondages, Ed. Univ. Bruxelles
- DROESBEKE J., Fichet B., Tassi P. (ed) (1987), Les Sondages, Economica
- DUSSAIX A., GROSBRAS J. (1993), Les sondages: Principes et methodes, Ed. Que Sais-Je?
- FODDY W. (1993), Como Perguntar, Celta
- GHIGLIONE R., MATALON B. (1993), O Inquérito, Celta
- GOURIEROUX C. (1981), Theorie des sondages, Economica

- GRANJÉ D., LEBART L. (1994), Traitement Statistiques des enquêtes, Dunod
- GROSBRAS J. (1987), Methodes Statistiques des Sondages, Economica
- GROSBRAS, J.-M. (1987), *Méthodes statistiques des sondages*, Paris, Economica.
- IDEA (1997), *Voter Turnout from 1945 to 1997 A Global Report.* Estocolmo, IDEA;
- JAVEAU C. (1992), L'enquête par questionnaire, Ed. Univ. Bruxelles
- JUSTEL M. (1995), La abstención electoral en Espanã (1977 - 1993), CIS
- KISH L. (1965), Survey Sampling, Wiley
- LEBART L. (ed) (1992), La qualité de l'information dans les enquêtes, Dunod
- LEHTONEU R., PAHKINEN E. (1995), Practical methods for design and analysis of complex surveys, Wiley
- LEVY P., LEMESHOW S. (1991), Sampling of populations: Methods and applications, Wiley
- MENDES, Maria de Fátima Abrantes e MIGUÉIS, Jorge - *Lei Eleitoral da Assembleia da Republica actualizada, anotada e comentada*;
- OLIVEIRA, Tiago de, (1981) - "O Sistema Eleitoral Português como Forma de Representação", *Análise Social*, vol. XVII (65);
- SARMENTO I. (1994), Tópicos da Teoria de amostragem, Ed. Univ. Trás-os-Montes e Alto Douro
- SÄRNDAL, C-E, Swensson, B., Wretman, J. (1992) *Model Assisted Survey Sampling*, New York, Springer-Verlag.
- STAPE - *Eleição da Assembleia da Republica 1995 (escrutínio provisório).*
- THOMPSON S. (1992), Sampling, Wiley

• TRYFON P. (1996), Sampling methods for applied research, Wiley
• VICENTE P., REIS E., FERRÃO F. (1996), Sondagens, Ed. Silabo

Comissão Organizadora Local do VI Congresso Anual da SPE

Daniel Paulino (*Instituto Superior Técnico*)

Ana Pires (*Instituto Superior Técnico*)

António Pacheco (*Instituto Superior Técnico*)

Ferreira da Cunha (*Instituto Nacional de Estatística*)

Amélia Leitão (*Direcção-Geral de Saúde*)

Apoios :

Câmara Municipal de Tomar

Centro de Estatística e Aplicações / UL

Centro de Matemática Aplicada / IST

EXPO'98 Lisboa

Instituto Nacional de Estatística

Instituto Superior de Estatística e Gestão de Informação

Instituto Superior Técnico

Região de Turismo dos Templários

Secção de Estatística e Aplicações / DM-IST

Livraria Escolar Editora

Timberlake Consultants