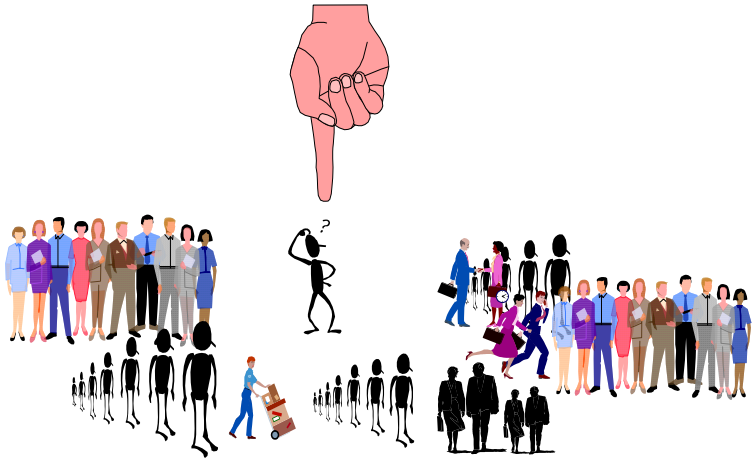
Instituto Superior de Economia e Gestão
UNIVERSIDADE TÉCNICA DE LISBOA

Sondagens

Notas para as aulas

Filomena G. Pimenta



Setembro de 2010

ÍNDICE

1. NOÇÕES BÁSICAS DE AMOSTRAGEM

Leitura obrigatória: capítulo 1 do livro "Sampling: Design and Analysis", Sharon L. Lohr

A recolha de informação relativamente a uma população pode ser efectuada quer de forma exaustiva, observando toda a população (censo), quer analisando somente uma fracção da população (inquérito por amostragem).

Dispor de informação sobre os fenómenos é essencial não só para o seu conhecimento mas, fundamentalmente, para uma correcta intervenção sobre a sua evolução, através da avaliação, de uma forma adequada e com base nos dados disponíveis, dos cenários equacionados no processo de tomada de decisão. Obter a informação necessária no momento certo e ao menor custo constitui o principal objectivo da Amostragem.

Um inquérito por amostragem, ou sondagem, é assim um inquérito levado a efeito sobre uma fracção da população estudada, designada por amostra. Estes estudos só terão interesse se, com base na informação recolhida na amostra, for possível estimar a distribuição das características da população.

1.1. Vantagens da amostragem

Se em princípio a observação exaustiva proporciona resultados exactos, ao contrário da amostragem que apenas proporciona resultados aproximados e não 100% seguros, pode perguntar-se qual a razão de usar a amostragem como técnica corrente, isto é, quais as suas vantagens. Enumerem-se, então, algumas das principais vantagens da amostragem:

- em primeiro lugar, permite redução dos custos e maior rapidez no apuramento dos resultados. De facto, se o número de elementos da população for elevado, como acontece correntemente, o método de observação exaustiva é dispendioso e moroso. Assim o estudo de um subconjunto da população (amostra) reduz substancialmente os custos e o tempo dispendido.
- em segundo lugar, permite uma maior profundidade no tratamento da informação. Com efeito, a utilização de grupos menores facilita a recolha de um maior número de dados sobre cada um dos elementos que compõem a amostra, possibilitando, deste modo, o aprofundamento de certos aspectos.
- em terceiro lugar, há que considerar que em certos casos a observação é destrutiva, isto é, implica a destruição dos elementos analisados. O exemplo clássico é o do controlo de qualidade da produção diária de fósforos. Se quisermos saber, por exemplo, a proporção de fósforos produzidos diariamente que não acendem, a observação exaustiva implica a necessidade de os acender a todos...
- finalmente, convém não esquecer que embora teoricamente a observação exaustiva conduza a resultados exactos, há sempre os erros de recolha da informação, que acontecem tanto na observação e anotação dos resultados como no próprio tratamento dos mesmos. É evidente que, ao recolher e processar uma menor quantidade de informação, diminuem as possibilidades deste tipo de erro principalmente porque se podem utilizar meios de recolha mais adequados e efectuar uma melhor supervisão do processo.

1.2. Plano de amostragem

Para levar a efeito um estudo por amostragem convém elaborar um *plano de amostragem*, isto é, definir claramente as etapas a percorrer, das quais salientamos as principais:

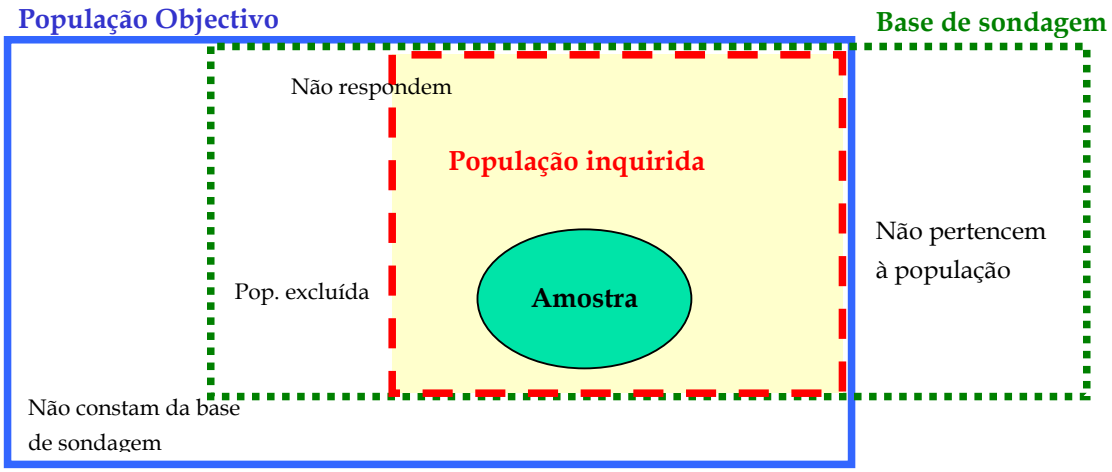
a) Definição dos objectivos do estudo

Nesta primeira etapa procede-se à avaliação das necessidades que o estudo visa satisfazer, precisando a natureza dos resultados a apurar, assim como os vários critérios, qualitativos e quantitativos, a ter em atenção quando da recolha de informação.

b) Definição da População

Esta fase está, como é evidente, fortemente ligada à anterior. *População objectivo* ou universo é o conjunto de todos os elementos cujas características queremos estudar. Cada um dos elementos que formam a população, chama-se unidade final (UF). Às vezes, por motivos de ordem prática, não é possível seleccionar a amostra a partir da população objectivo. Nestes casos torna-se necessário redefinir uma população que esteja estreitamente relacionada com a primeira, designada por *população inquirida* que é o conjunto de todas as UF que podem vir a ser seleccionadas para integrar a amostra, servirá assim, de base à amostragem. Convém ter presente que as inferências feitas com base num estudo por amostragem são estatisticamente válidas para esta população, população inquirida. Assim o ideal será que a população objectivo e a população inquirida coincidam. A definição da população deve ser feita sem ambiguidade de forma a facilitar o trabalho de recolha da informação. Algumas definições de população reflectem facilidade na identificação das unidades básicas, tais como: clientes em atraso ou estudantes universitários. O mesmo não acontece com outras populações que não são facilmente identificadas só pela designação como, por exemplo, médias empresas ou classe média. As ambiguidades devem ser previstas e a sua resolução faz-se normalmente através de convenções que clarifiquem as situações dúbias.

O ideal será definir a matriz da população ou *base de sondagem*, listagem das unidades amostrais de onde será seleccionada a amostra. *Unidade amostral* (UA), unidade que é seleccionada para integrar a amostra. Pode não coincidir com a unidade final, por exemplo, pode querer-se inquirir pessoas, UF = indivíduo, e ter-se seleccionado uma amostra de agregados familiares, ou seja, UA = Família.



c) Identificação, selecção e análise da informação existente

Com o objectivo de caracterizar minimamente o fenómeno em estudo e permitir a escolha do método de análise mais adequado, é importante estudar a informação existente considerada relevante, nomeadamente estatísticas publicadas e inquéritos análogos ou complementares. Esta informação, informação secundária (interna ou externa à organização), existe em quantidade que é geralmente desconhecida dos pesquisadores, e é um precioso auxiliar tanto na definição do problema e formulação de hipóteses como na decisão de qual a natureza e a forma da recolha dos dados de um projecto de pesquisa.

d) Escolha do método de amostragem

Esta fase é crucial e depende de múltiplos factores, nomeadamente, de restrições materiais (meios financeiros e humanos afectos, duração do inquérito, meios informáticos existentes...), dos resultados da análise da informação disponível e dos objectivos a atingir.

Os métodos de amostragem podem ser classificados em dois grandes grupos: *métodos aleatórios ou probabilísticos* e *métodos não aleatórios ou empíricos*. Só os métodos probabilísticos é que associam uma variância a cada estimador e possibilitam a determinação da sua distribuição de probabilidade, permitindo quantificar o erro de amostragem que decorre da não observação de toda a população. Nos métodos de amostragem empíricos não é possível essa quantificação. No entanto, são muito utilizados, em particular nas sondagens de opinião, pois apresentam um certo número de vantagens económicas: menores custos, pessoal reduzido, obtenção rápida de resultados.

e) Quantificação da amostra

Ao planear um levantamento por amostragem é fundamental a determinação da dimensão da amostra ($\boldsymbol{n}$). Uma amostra demasiado grande implica um desperdício de recursos, e uma muito pequena diminui a credibilidade dos resultados.

Então qual deverá ser a dimensão da amostra?

Esta questão simples e incisiva não tem uma resposta imediata. A Estatística proporciona um corpo teórico e os instrumentos necessários para apreciar e responder a esta questão depois de reequacioná-la. A dimensão da amostra depende basicamente de quatro factores:

- número de grupos e subgrupos a analisar;
- valor da informação do estudo, em particular, nível de precisão e grau de confiança pretendidos para os resultados, isto é, o erro de amostragem máximo admitido e a sua probabilidade de ocorrer;
- custo de obtenção da amostra e orçamento disponível;
- variabilidade da característica na população: se os elementos da população têm comportamentos similares, a dimensão da amostra pode ser pequena, enquanto que para uma população muito heterogénea a amostra deverá ser maior.

Esclareça-se, desde já, que a designação correntemente utilizada de taxa de amostragem (f) representa, em populações finitas, o quociente entre a dimensão da amostra e a dimensão da população ($f = n/N$).

f) Escolha do método de recolha de dados

Existem basicamente três métodos de recolha de dados: observação, experimentação e levantamento. O levantamento é, sem dúvida alguma, o método mais popular de recolha de dados que envolvam comportamentos pessoais ou organizacionais. O método define-se pela recolha da informação com base numa amostra representativa através de comunicação pessoal. Colocam-se às pessoas, ou organizações, questões que deverão ser respondidas oralmente ou por escrito.

Os instrumentos básicos são o questionário e a entrevista. Os levantamentos são o principal meio de recolha de informação primária dos estudos de mercado. Existem vários tipos de levantamento que se distinguem, sobretudo pela forma como se comunica com os entrevistados. Os principais tipos são: entrevista pessoal pré-seleccionada, entrevista pessoal de interceptação, entrevista telefónica, entrevista em profundidade, questionários enviados pelo correio, questionários entregues aos respondentes para preenchimento sendo posteriormente apanhados pelos entrevistadores, painéis e entrevistas em grupo. Todos os tipos têm vantagens e inconvenientes e a escolha de um deles terá que ver com os objectivos do estudo e orçamento disponível.

g) Recolha, codificação, verificação, análise e interpretação da informação

Antes de levar a efeito um inquérito por amostragem, é necessário definir com clareza um conjunto de elementos. A ausência de tal definição pode conduzir a erros de recolha impossíveis de quantificar, afectando a amostra e os resultados do estudo, pois podem ser mais importantes que os erros de amostragem. A qualidade dos dados a obter é um requisito fundamental sem o qual não é possível a correcta extrapolação dos resultados. Para garantir essa qualidade, é fundamental a preparação do instrumento de notação e uma boa organização do trabalho de campo.

O instrumento de notação (questionário ou guia de entrevista) deve ser elaborado, de forma a traduzir os objectivos específicos do estudo numa linguagem que seja acessível às unidades estatísticas que compõem a amostra. A relevância, eficácia, sequência e forma de apresentação das perguntas devem ser cuidadosamente estudadas e ensaiadas mediante a realização de um pré-teste ao questionário.

Deve ainda estabelecer-se quais os dados que necessitam de codificação e qual a codificação, de forma a possibilitar o tratamento informático da informação e simultaneamente definir os procedimentos de controlo de qualidade desses dados com o objectivo de reparar erros de medida que possam acontecer na recolha.

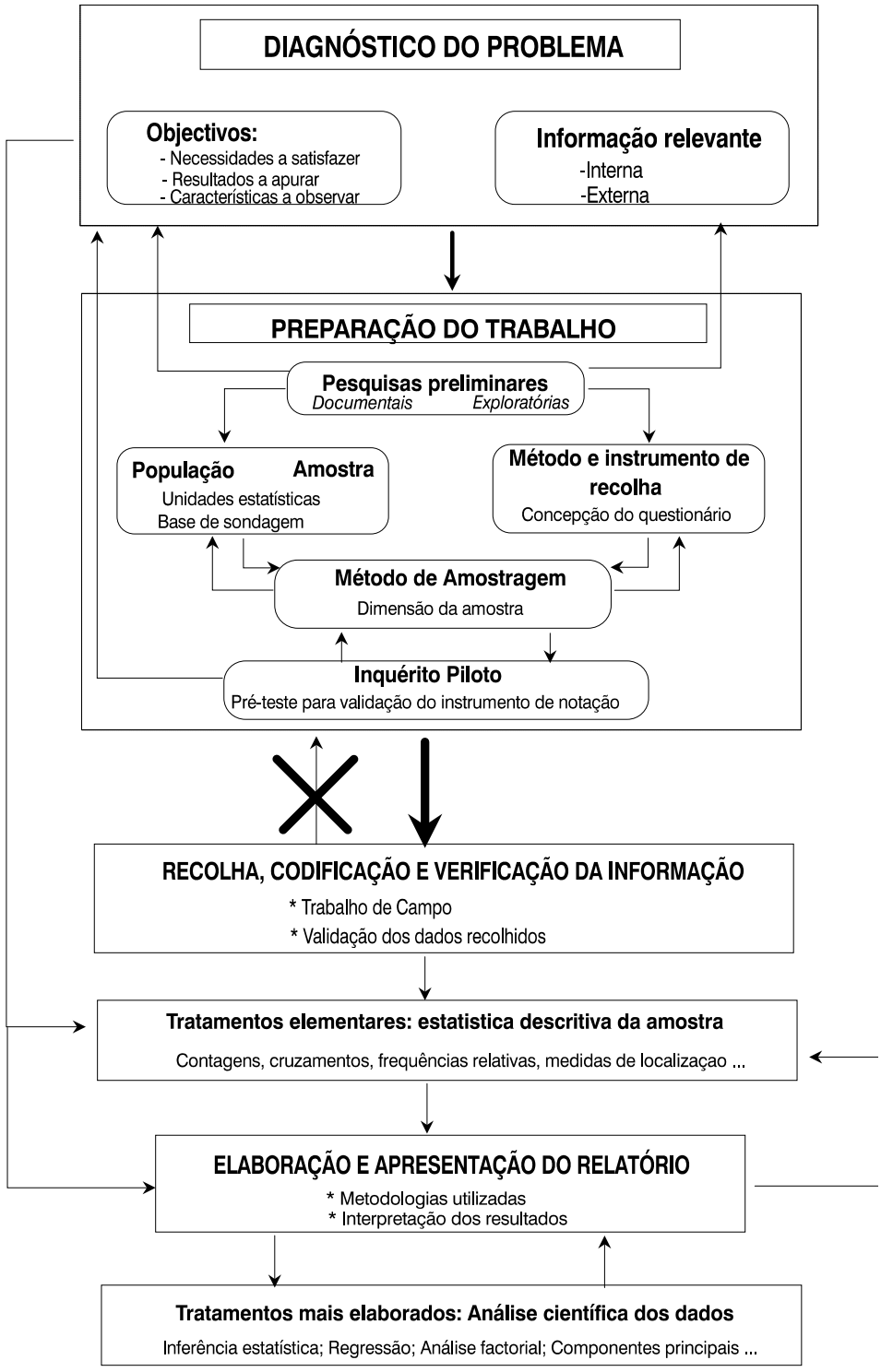
É também importante a definição de quais os procedimentos estatísticos a utilizar tanto para a análise preliminar dos dados como para a obtenção dos resultados finais.

h) Apresentação dos resultados

O relatório final do trabalho deverá:

- resumir as fases principais da preparação e execução do levantamento;
- indicar e justificar tanto as hipóteses subjacentes como os métodos e técnicas utilizadas.
- apresentar de uma forma clara a informação recolhida, por exemplo, com quadros resumo e gráficos correctos, e comentá-la, formulando hipóteses e conclusões que possam ser sustentadas por esses mesmos dados.

FASES DE UM INQUÉRITO POR AMOSTRAGEM



1.3. Erros de amostragem e erros de recolha

Os erros de amostragem ocorrem porque uma amostra não é uma miniatura perfeita da população. Por exemplo, a média aritmética dos valores observados na amostra não coincidirá, necessariamente, com o verdadeiro valor da média da população, será uma estimativa. No entanto, se a amostragem for aleatória é possível, como já foi referido, quantificar esse erro e tentar mantê-lo dentro de certos limites, nomeadamente, através de variações na dimensão da amostra.

Os erros de recolha de dados não são claramente definidos, existem independentemente do tipo de amostragem, dificilmente são medidos ou estimados e contrariamente aos erros de amostragem tendem a aumentar com a dimensão da amostra. Podem acontecer em todas as fases da sondagem, e são susceptíveis de afectar a validade dos resultados.

São principalmente três as fontes de erro:

- cobertura inadequada da população,
- erros na recolha e codificação da informação,
- falta de resposta de algumas unidades estatísticas.

Diz-se que a cobertura da população é inadequada quando elementos representativos deixam de ser incluídos na amostra. A principal razão para a ocorrência deste problema está directamente ligada à matriz da população. Grande parte das bases de sondagem são listagens incompletas, desactualizadas ou distorcidas, que não garantem a réplica da população. Muitos pesquisadores utilizam listas que não contêm certos segmentos representativos do universo em análise. Por exemplo, utilizar como base a lista telefónica exclui grande parte das famílias de baixos rendimentos. É pois importante, detectar falhas na base de sondagem procurando, se possível, cruzar várias fontes de informação.

Os erros na recolha e codificação de informação: podem resultar de má preparação do instrumento de notação, falta de controlo da qualidade do trabalho de campo e interpretações subjectivas erróneas das respostas dadas, feitas pelo entrevistador ou pelo codificador.

A expressão "falta de resposta" designa a impossibilidade de obtenção de resposta dos elementos da amostra e acontece quer por não se conseguirem contactar as unidades estatísticas, quer por sua recusa em responderem aos questionários. Este problema é um quebra-cabeças para o investigador, sobretudo quando as taxas de não resposta são elevadas, porque se não for tido em linha de conta pode provocar enviesamentos significativos nos resultados.

Temos basicamente três situações para não resposta:

- Falhas na base de sondagem – Impossibilidade da localização ou da visita a algumas unidades da amostra decorrente normalmente de bases de sondagem desactualizadas. Neste caso, é conveniente ou ter, paralelamente à amostra principal, amostras de substituição recolhidas com a mesma metodologia da amostra principal, ou estabelecer claramente regras de substituição.
- Não encontrados – Este grupo contém as pessoas que não se encontram temporariamente no local previsto. Convém definir regras ou políticas de entrevistas para evitar esse problema. Em geral, programam-se novos contactos em horários diferentes. Todos sabemos que as famílias em que ambos os cônjuges trabalham fora, ou

que não têm crianças, são mais difíceis de encontrar do que as famílias que têm crianças de pouca idade ou pessoas idosas.

- Recusa de participação – A unidade estatística existe, foi contactada, mas não responde quer por negligência quer por recusa. Neste caso deverá ser elaborada uma metodologia de tratamento que permita a modelação do comportamento desta fracção da população, pois podemos cometer erros considerando comportamento similar aos que responderam.

A forma de tratamento deste problema dependerá da análise feita a dados e informações já existentes. Normalmente fazem-se insistências junto a uma subamostra dos que não responderam, tentando obter o mínimo de informação sobre esta parte da população ou utilizam-se técnicas de imputação das não respostas ou de reajustamento amostral.

1.4. Métodos de amostragem

Como já referido, distinguem-se duas grandes categorias de amostras: as amostras aleatórias ou probabilísticas e as amostras não aleatórias.

A amostragem é aleatória quando cada elemento da população tem uma probabilidade conhecida e diferente de zero de integrar a amostra, designada por probabilidade de inclusão. Tendo-se como pressuposto que as variáveis observadas na amostra são aleatórias, a partir delas é possível tanto estimar as grandezas correspondentes da população como quantificar o erro associado a essa estimação (erro de amostragem).

A amostragem é não aleatória, quando as probabilidades dos elementos da população pertencerem à amostra são desconhecidas. Neste tipo de amostragem a construção da amostra é feita a partir de informações *a priori* sobre a população estudada, tentando que a amostra seja um espelho tão fiel quanto possível dessa população. Trata-se de um conjunto de métodos empíricos envolvendo juízos de valor de quem as selecciona, não permitindo, por isso, avaliar a precisão das estimativas obtidas.

1.4.1. Amostragem não aleatória: métodos empíricos

As amostras não aleatórias ou não probabilísticas são classificadas em vários tipos dos quais se destacam:

- orientadas ou por julgamento,
- arbitrárias ou por conveniência,
- amostras "bola de neve"
- amostragem por quotas.

A) Amostragem orientada

Quando a selecção dos elementos da população é feita por especialistas, em função das propriedades que possuem relativamente aos objectivos da pesquisa. É um procedimento comum na investigação exploratória, em que são seleccionados peritos em determinados temas, para serem entrevistados.

Um exemplo deste tipo de amostragem é a obtenção de previsões de resultados eleitorais nacionais, com base em certas freguesias, previamente seleccionadas, consideradas pelos especialistas como representativas da população.

B) Amostragem por conveniência

Como o próprio nome indica, a selecção é feita de forma arbitrária em função da conveniência da pesquisa (uma turma de estudantes, os clientes dum centro comercial num determinado dia, questionários incluídos num semanário).

C) Amostras "bola de neve"

Este método é utilizado sobretudo quando a população a analisar é constituída por casos dificilmente encontrados. Procede-se inicialmente a um inquérito sobre um nº elevado de pessoas, para apurar as que pertencem à população, isto é, apresentam a característica rara. Interrogam-se estas pessoas sobre os problemas cobertos pelo inquérito e pede-se-lhes a indicação de outras que também tenham essa característica de excepção. Assim se vai construindo a amostra, até atingir a dimensão pretendida.

D) Amostragem por quotas

O método de amostragem por quotas, usado frequentemente em estudos de mercado e inquéritos de opinião, baseia-se na hipótese da existência de correlação entre as diferentes características duma população. Sendo válida esta hipótese, será natural que numa amostra escolhida de modo a ter distribuição de certas características (variáveis de controlo) idêntica à da população da qual é retirada, também se verifique esse comportamento relativamente aos aspectos em estudo. A amostra assim construída possui, para as variáveis de controlo, a mesma estrutura da população. As quotas, a serem respeitadas pelos entrevistadores, obtêm-se multiplicando a taxa de amostragem pelo número de elementos da população pertencentes a cada classe definida pelas variáveis de controlo.

As variáveis de controlo mais usadas são as provenientes dos dados demográficos ou sobre a actividade económica tais como: áreas geográficas, sexo, idade, habilitações académicas, escalões de rendimento, número de empregados, sector de actividade e volume de vendas.

A escolha dos elementos que compõem a amostra é, geralmente, da exclusiva responsabilidade do entrevistador, só tendo que respeitar as quotas fixadas.

1.4.2. Amostragem aleatória: métodos probabilísticos

Existem vários tipos de amostragem aleatória, dos quais destacamos a amostragem:

- aleatória simples
- estratificada
- por conglomerados ou agrupamento
- sistemática
- por etapas múltiplas ou multi-etápica e
- por fases múltiplas.

A) Amostragem aleatória simples

Uma amostra diz-se aleatória simples, quando cada elemento do universo tem a mesma probabilidade de ser escolhido para entrar na amostra. Neste processo de escolha deixa-se inteiramente ao acaso a indicação de quais os elementos a incluir na amostra. A principal dificuldade da escolha aleatória simples é de ordem prática e surge imediatamente ao pôr-se a seguinte questão: como obter uma amostra aleatória simples?

A forma mais comum consiste na atribuição de uma sequência de números aos elementos da população, e em seguida gerar números aleatoriamente, tantos quantos a dimensão da amostra o exigir. Como cada número corresponde a um elemento da população, obtém-se a amostra fazendo a correspondência dos números. Os números aleatórios podem ser obtidos gerando-os em computador, utilizando tabelas de números aleatórios publicadas em livros de Estatística e similares ou ainda sorteando números de uma urna que contenha a sequência de números correspondente aos elementos da população (método da lotaria).

A selecção pode fazer-se *com reposição* ou *sem reposição*, isto é, uma vez escolhido um número podemos ou não admitir que ele possa ser sorteado novamente. Quando a extracção é feita com reposição, pode suceder que uma mesma unidade estatística venha a figurar mais que uma vez na amostra, o que nunca sucede quando a escolha se faz sem reposição.

Normalmente, no caso da amostragem aleatória simples a partir de universos finitos, a escolha é feita sem reposição.

Se a selecção é feita com reposição as sucessivas extracções são independentes, podendo extrair-se N^n amostras equiprováveis (cada uma delas com probabilidade $1/N^n$) onde $\boldsymbol{N}$ é o número de elementos da população e $\boldsymbol{n}$ a dimensão da amostra. Se for sem reposição as extracções não são independentes e o número de amostras equiprováveis é de C_n^N .

Quando a dimensão da amostra não é significativa em relação à dimensão do universo, como correntemente sucede, não há diferenças significativas entre a escolha aleatória simples sem reposição e com reposição.

A amostragem aleatória simples com reposição, designada correntemente por amostragem casual (onde as $\boldsymbol{n}$ variáveis aleatórias observadas $X_1, X_2, ..., X_n$ são independentes e identicamente distribuídas), é o pressuposto de todos os resultados de inferência estatística desenvolvidos em qualquer manual de Estatística.

B) Amostragem Estratificada

Na amostragem estratificada a população é subdividida em classes mais homogéneas denominadas estratos, de cada um dos quais se seleccionam amostras aleatórias que combinadas formam a amostra total. Para subdividir a população em estratos é indispensável a existência de informação suplementar sobre alguma ou algumas características da população (variáveis de estratificação).

Para que uma característica, quantitativa ou qualitativa, seja escolhida como variável de estratificação deverá:

- estar correlacionada com as variáveis em estudo;
- ter um valor previamente conhecido para cada um dos elementos da população.

Se a característica não for conhecida com precisão, os erros de classificação cometidos na constituição dos estratos podem, reduzindo a sua homogeneidade, diminuir a eficácia do método.

A estratificação é muito utilizada pois possui várias vantagens de que se destacam três:

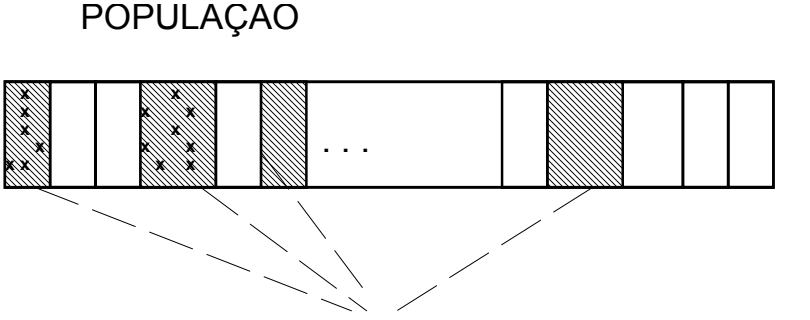
- permite a obtenção de estimativas com determinada precisão para certas subdivisões da população (domínios de estudo), sendo aconselhável tratar cada uma das subdivisões como uma população;
- proporciona um aumento de precisão nas estimativas das características da totalidade da população, sem aumentos significativos nos custos. Esse aumento de precisão será tanto maior quanto mais homogéneas forem as unidades estatísticas dentro de cada estrato e mais heterogéneos os estratos;
- conveniências administrativas de organização do trabalho de recolha da informação. Por exemplo, a entidade que realiza o estudo pode ter departamentos ou estruturas regionais que podem levar a cabo o trabalho de campo de uma parte da população.

C) Amostragem por conglomerados (one-stage cluster sampling)

Na amostragem por conglomerados, também chamada por cachos, as unidades estatísticas são agrupadas, de acordo com algum critério, em conjuntos chamados conglomerados pertencendo cada unidade estatística a um e um só deles. Por exemplo, um estabelecimento comercial constitui um conglomerado de empregados de comércio.

A amostragem por conglomerados diferencia-se da aleatória simples, porque as unidades estatísticas que compõem a amostra não são escolhidas uma a uma, mas em grupos (conglomerados). Estes são seleccionados aleatoriamente, sendo observadas todas as unidades que o compõem.

Este tipo de amostragem é utilizado quando não se conhece bem a composição da população, ou quando, principalmente por uma questão de recursos, se torna necessário limitar as áreas de investigação.



Conglomerados da amostra: serão observadas todas as unidades est
que fazem parte destes conglomerados

O principal óbice deste tipo de amostragem está em poder acontecer que unidades estatísticas pertencentes a um mesmo conglomerado tendam a assemelhar-se para certas características. Por exemplo, um prédio de habitação, pode ser um conglomerado relativamente eficaz para estimar a repartição de uma população relativamente a sexo, idade, ocupação, mas talvez já não o seja para um estudo por classes de rendimento pois neste caso pode acontecer que as unidades sejam mais homogéneas.

Quando as unidades estatísticas se agrupam de acordo com a sua proximidade geográfica é designada por amostragem por áreas.

D) Amostragem sistemática

Existe um outro tipo de amostragem que se pode apresentar como uma particularização da amostragem por conglomerados e que merece especial destaque: a amostragem sistemática. A amostragem sistemática corresponde à escolha de um único conglomerado constituído por todas as unidades estatísticas cujos números de referência façam parte da mesma progressão aritmética.

Para se extrair uma amostra sistemática, deve-se obter uma lista sequencial dos elementos da população. Calcula-se depois o intervalo amostral ($I_k=N/n$, que por comodidade supomos um número inteiro) que é a razão entre a dimensão da população e a dimensão da amostra.

Em seguida selecciona-se aleatoriamente um número entre um e o intervalo amostral, número de arranque (designado por ***a***), que serve tanto para determinar o ponto de partida na lista sequencial como para indicar o primeiro elemento a integrar a amostra.

A este número, soma-se sucessivamente o intervalo amostral e, os elementos correspondentes às ordens dos números obtidos (progressão aritmética de razão igual ao intervalo amostral) serão os outros elementos da amostra. Assim, a amostra será constituída pelas ***n*** unidades estatísticas a que correspondem os números:

$$a \ , \ a+I_k \ , \ a+2I_k \ , \ \dots \ , \ a+(n-1)I_k$$

O uso da amostragem sistemática é comum, principalmente pela sua facilidade, pois só requer a geração de um número aleatório de arranque. No entanto deve-se tomar cuidado pois a composição das listas pode afectar a representatividade da amostra.

Explicitando melhor, se na lista sequencial:

- a ordem das unidades for considerada aleatória, este tipo de amostragem é equivalente à aleatória simples sem reposição;
- os elementos consecutivos têm comportamentos similares, a amostragem sistemática garantirá maior precisão que a aleatória simples;
- as unidades tiverem sido ordenadas com um critério de que resulte o aparecimento periódico de unidades estatísticas com comportamentos similares, a amostragem sistemática poderá conduzir a graves erros de estimação, sobretudo se o período for um submúltiplo da razão da progressão aritmética.

E) Amostragem multi-etápica

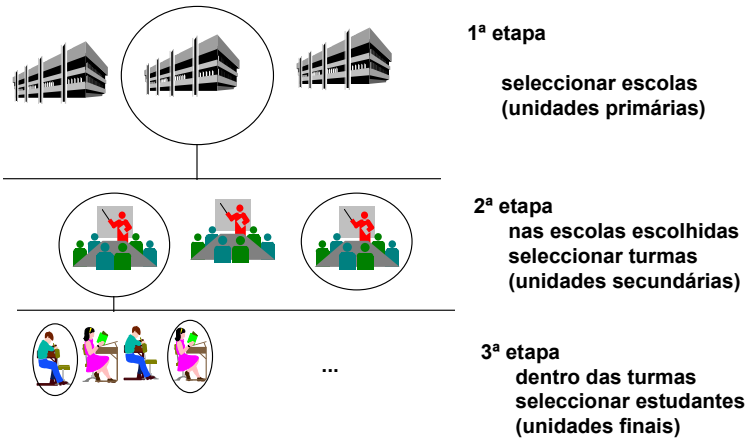
Uma outra técnica de amostragem bastante aplicada é a amostragem multi-etápica que consiste em estabelecer uma hierarquia de unidades de amostragem - unidades primárias, secundárias, terciárias etc. - e a escolha das unidades finais a observar é feita não num estágio único, mas em várias etapas (como o próprio nome indica).

Assim:

- na 1ª etapa é escolhida uma amostra aleatória de unidades primárias;
- na 2ª etapa retira-se, de cada unidade primária da amostra da 1ª etapa, uma amostra aleatória de unidades secundárias;
- na 3ª etapa retira-se, em cada unidade secundária da amostra anterior, uma amostra aleatória de unidades terciárias;
- e etc.

EXEMPLO:

A selecção de uma amostra aleatória de estudantes pode ser feita segundo o esquema seguinte:



Quando o número de etapas é de duas ou três a amostragem chama-se, respectivamente, bi-etápica (two-stage cluster sampling) ou tri-etápica.

EXEMPLO:

Suponhamos que queremos seleccionar uma amostra aleatória de explorações agrícolas. Em vez de listar exaustivamente as explorações agrícolas de todo o país e retirar uma amostra aleatória simples, poderíamos:

- primeiro escolher aleatoriamente os distritos que seriam as unidades primárias;
- na 2ª etapa seleccionar, em cada um dos distritos que compõem a amostra anterior, uma amostra aleatória de concelhos (unidades secundárias);
- por fim, 3ª etapa, para cada concelho escolhido, obter uma listagem completa das explorações agrícolas existentes, e retirar uma amostra aleatória destas explorações (unidades terciárias = unidades finais).

De entre as vantagens da amostragem multi-etápica salientamos:

- simplificação na definição da base da sondagem. Não é necessário construir toda a matriz da população: no exemplo acima, só precisamos conhecer a lista das explorações agrícolas dos concelhos escolhidos;
- para um mesmo número de unidades examinadas, diminuição significativa dos custos, pois normalmente verifica-se uma maior concentração geográfica das unidades a observar, o que permite reduzir as despesas de deslocação.

F) Amostragem por fases múltiplas

A amostragem por fases múltiplas emprega-se para, a partir de uma subamostra da amostra principal, se obterem informações mais pormenorizadas sobre determinados aspectos, análise que não convém alargar a todos os elementos que compõem a amostra inicial (ou população no caso de um censo) por ser difícil, pouco económico ou demorado o apuramento de resultados.

G) Outros métodos

Na prática um estudo por amostragem não utiliza normalmente um só tipo de amostragem mas sim uma combinação de vários métodos, que melhor garanta a prossecução dos objectivos, daí a designação de *sondagens complexas*.

Em todos os métodos de amostragem aleatórios referidos a dimensão da amostra é fixada antecipadamente. Mas existe outra classe de métodos de amostragem chamados *sequenciais* ou progressivos, onde a dimensão da amostra não é fixada *a priori*, mas a partir de regras que têm por base as próprias observações.

1.5. Enviesamento e variabilidade amostral

Construir ***estimadores*** - T_n – para o(s) parâmetro(s) desconhecido(s) - θ - e avaliar a sua "qualidade" estudando o enviesamento e a variabilidade amostral (assumindo a não existência de erros de medida).

Validade - capacidade do estimador em produzir estimativas centradas no verdadeiro valor do parâmetro a estimar. ***(validity)***

Medida pelo enviesamento : $Env(T_n) = E(T_n) - \theta$

Se $Env(T_n) = 0$, isto é, se $E(T_n) = \theta$ então T_n centrado ou não enviesado

Fiabilidade - capacidade do estimador em produzir estimativas próximas da sua média. ***(reliability)***

Avaliada pela variância: $V(T_n) = E\left[T_n - E(T_n)\right]^2$ ou desvio padrão

mede o grau de dispersão em torno da sua média

Precisão - capacidade do estimador em produzir estimativas próximas do verdadeiro valor do parâmetro a estimar ***(accuracy)***

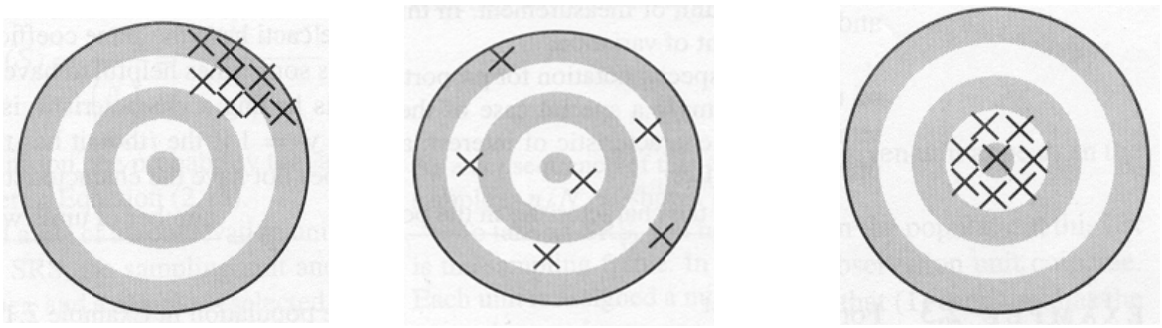
Medida pelo Erro Quadrático Médio (EQM ou MSE)

$$EQM(T_n) = E\left[(T_n - \theta)^2\right] = V(T_n) + \left[Env(T_n)\right]^2 \text{ ou } \sqrt{EQM(T_n)}$$

mede o grau de dispersão em torno do verdadeiro valor de θ

Se T_n centrado então $EQM(T_n) = V(T_n)$

A **B** **C**



Enviesado mas fiável Não enviesado mas pouco fiável Grande precisão

2. AMOSTRAGEM ALEATÓRIA SIMPLES

Leitura obrigatória: capítulo 2 do livro "Sampling: Design and Analysis", Sharon L. Lohr

2.1. Conceitos e notação

População composta por ***N*** elementos (população finita)

$U_1 \quad U_2 \quad ... \quad U_s \ldots \quad U_N$

cada um deles com intensidade da característica em estudo, respectivamente,

$x_1 \quad x_2 \quad ... \quad x_s \ldots \quad x_N$

Amostra composta por ***n*** elementos

$X_1 \quad X_2 \quad ... \quad X_i \ldots \quad X_n.$

Onde cada X_i é uma v. a. com distribuição dada por:

valores	x_1	x_2	...	x_s	...	x_N
probabilidades	*1/N*	*1/N*	...	*1/N*	...	*1/N*

quer a selecção seja feita com ou sem reposição

Notação

- **Para a população**

$t = \sum_{s=1}^{N} x_s$ [ ] notação correspondente no livro, não utilizada nos acetatos

$$[\bar{x}_U =] \mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{s=1}^{N} x_s}{N}$$

$$\sigma^2 = \frac{\sum_{s=1}^{N} (x_s - \mu)^2}{N} = \frac{\sum_{s=1}^{N} x_s^2}{N} - \mu^2 \qquad \text{ou} \qquad [S^2=]\ \sigma'^2 = \frac{\sum_{s=1}^{N} (x_s - \mu)^2}{N-1} = \frac{N\sigma^2}{N-1}$$

- **Para a amostra**

$$[\bar{x}_S =]\ \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n} = \frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}^2 \qquad [s^2=]\ S'^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{nS^2}{n-1}$$

⇒ **Selecção com reposição (PICR)**

- tiragens feitas rigorosamente nas mesmas condições, independentes
- em cada tiragem U_s tem probabilidade 1/N de ser seleccionada
- a probabilidade de U_s pertencer à amostra é igual a $1-(1-1/N)^n$

⇒ **Selecção sem reposição (PISR) - tiragens exaustivas**

- já não há independência
- em cada tiragem U_s tem probabilidade 1/N de ser seleccionada
- a probabilidade de U_s pertencer à amostra é igual a $\frac{C_1^1 C_{n-1}^{N-1}}{C_n^N} = \frac{n}{N}$

2.2. Estimação da média da população e erro padrão associado

$$\mu = \frac{\sum_{s=1}^{N} x_s}{N} \qquad \xrightarrow{estimador} \qquad \hat{\mu} = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Estimador centrado ou não enviesado: $E(\overline{X}) = \mu$

- **Variância deste estimador**

PICR		***PISR***
$\sigma^2_{\overline{X}} = V(\overline{X}) = \frac{\sigma^2}{n}$	=	$\frac{N-n}{N-1}\frac{\sigma^2}{n} = (1-\frac{n}{N})\frac{\sigma'^2}{n}$ (ver demonstração na página seguinte)

$\Downarrow$

coeficiente de exaustividade ou c.p.f.
(correcção de populações finitas)

$$\boxed{V_{PICR}(\overline{X}) \geq V_{PISR}(\overline{X})}$$ ***Maior precisão PISR***

- **Desvio padrão do estimador**

	PICR		***PISR***
$\sigma_{\overline{X}} = \sqrt{V(\overline{X})} =$	$\frac{\sigma}{\sqrt{n}}$	=	$\sqrt{(1-\frac{n}{N})}\frac{\sigma'}{\sqrt{n}}$

Dedução da $V(\overline{X}) = \frac{N\text{-}n}{N\text{-}1}\frac{\sigma^2}{n} = (1-\frac{n}{N})\frac{\sigma'^2}{n}$ **se PISR**

$$V(\overline{X}) = V\left(\frac{\sum X_i}{n}\right) = \frac{1}{n^2}\left[\sum V(X_i) + \sum_{i\neq j}\sum Cov(X_i, X_j)\right]$$

$$\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) = E\left[\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}(X_i-\mu)(X_j-\mu)\right] = E\left[\sum_{\substack{\alpha=1\\ \alpha\neq\beta}}^{N}\sum_{\beta=1}^{N}(x_\alpha-\mu)(x_\beta-\mu)Z_\alpha Z_\beta\right] =$$

$$\sum_{\substack{\alpha=1\\ \alpha\neq\beta}}^{N}\sum_{\beta=1}^{N}(x_\alpha-\mu)(x_\beta-\mu)E(Z_\alpha Z_\beta) = \sum_{\substack{\alpha=1\\ \alpha\neq\beta}}^{N}\sum_{\beta=1}^{N}(x_\alpha-\mu)(x_\beta-\mu)\frac{n(n-1)}{N(N-1)}$$

Onde, $Z_\alpha = \begin{cases}1 & se\ U_\alpha \in S\\ 0 & se\ U_\alpha \notin S\end{cases}$ e

$$Z_\beta = \begin{cases}1 & se\ U_\beta \in S\\ 0 & se\ U_\beta \notin S\end{cases}$$

		Z_β	
		0	**1**
Z_α	**0**	$\frac{C_n^{N-2}}{C_n^N}$	$\frac{C_1^1 C_{n-1}^{N-2}}{C_n^N}$
	1	$\frac{C_1^1 C_{n-1}^{N-2}}{C_n^N}$	$\frac{C_{n-2}^{N-2}}{C_n^N} = \frac{n(n-1)}{N(N-1)}$

Como $\left[\sum_{\alpha=1}^{N}(x_\alpha-\mu)\right]^2 = 0 \Leftrightarrow \sum(x_\alpha-\mu)^2 + \sum_{\substack{\alpha=1\\ \alpha\neq\beta}}^{N}\sum_{\beta=1}^{N}(x_\alpha-\mu)(x_\beta-\mu) = 0$

$$\sum_{\substack{\alpha=1\\ \alpha\neq\beta}}^{N}\sum_{\beta=1}^{N}(x_\alpha-\mu)(x_\beta-\mu) = -\sum(x_\alpha-\mu)^2$$

$$\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) = -\frac{n(n-1)}{N(N-1)}\sum_{\alpha=1}^{N}(x_\alpha-\mu)^2 = -\frac{n(n-1)}{N(N-1)}N\sigma^2$$

Assim

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

⇒ **Erro padrão do estimador. Precisão amostral**

Erro Padrão de T_n $\quad \boxed{EP(T_n) = \sqrt{E\hat{Q}M(T_n)}}$

Se T_n não enviesado então $EP(T_n) = \sqrt{\hat{V}(T_n)}$

Para medir a ***precisão absoluta*** de T_n utiliza-se o ***E**rro **P**adrão do estimador*. *Quanto* ***menor o erro padrão*** *do estimador* ***maior a sua precisão.***

Para medir a ***precisão relativa*** de T_n utiliza-se Erro Relativo de Amostragem (ERA)

$$C\hat{V}(T_n) = \frac{EP(T_n)}{T_n}$$

- **Estimador para** σ^2

Poder-se-á utilizar a variância corrigida da amostra? S'^2

	PICR	PISR
$E(S'^2) =$	σ^2	$= \frac{N}{N-1}\sigma^2 = \sigma'^2$

então

$\hat{\sigma}^2_{\overline{X}} = \hat{V}(\overline{X}) =$	$\frac{S'^2}{n}$	$= (1 - \frac{n}{N})\ \frac{S'^2}{n}$

	PICR	**PISR**
$EP(\overline{X})$	$\frac{s'}{\sqrt{n}}$	$\sqrt{1-\frac{n}{N}}\ \frac{s'}{\sqrt{n}}$

2.3. Estimação de um total e de uma diferença

2.3.1. Estimador de um total

$$t = \sum_{s=1}^{N} x_s = N\mu \qquad \text{estimador} \qquad \hat{T} = N\overline{X} \qquad \text{não enviesado}$$

com

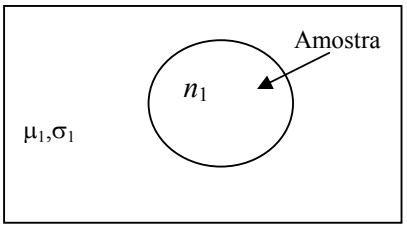
$V(\hat{T}) =$	$\frac{N^2\sigma^2}{n}$	***PICR***	$\hat{V}(\hat{T}) = \frac{N^2 S'^2}{n}$
$=$	$N(N-n)\ \frac{\sigma'^2}{n}$	***PISR***	$= N(N-n)\ \frac{S'^2}{n}$

2.3.2. Estimador de uma diferença

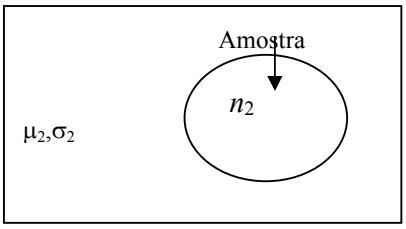
Comparar duas populações segundo o valor dum parâmetro (por exemplo as médias)

1º - Amostras independentes

População 1 - N_1 elementos



População 2 - N_2 elementos



Estimar $D = \mu_1 - \mu_2$ estimador $\hat{D} = \bar{X}_1 - \bar{X}_2$ não enviesado

Com variância $\sigma^2_{\hat{D}} = V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2)$

	PICR	***PISR***
$\sigma^2_{\hat{D}} =$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$	$= (1-\frac{n_1}{N_1})\ \frac{\sigma_1'^2}{n_1} + (1-\frac{n_2}{N_2})\ \frac{\sigma_2'^2}{n_2}$
$\hat{\sigma}^2_{\hat{D}} =$	$\frac{S_1'^2}{n_1} + \frac{S_2'^2}{n_2}$	$= (1-\frac{n_1}{N_1})\ \frac{S_1'^2}{n_1} + (1-\frac{n_2}{N_2})\ \frac{S_2'^2}{n_2}$

2º - Amostras emparelhadas

(Doentes observados antes e depois de tratamento; vendas antes e depois de campanha...)

Amostra 1	***População***	Amostra 2	
X_{11}	μ_1, σ_1 $\quad$ μ_2, σ_2	X_{21}	$D_1 = X_{11} - X_{21}$
X_{12}		X_{22}	$D_2 = X_{12} - X_{22}$
…		…	…
X_{1n}	σ_{12}	X_{2n}	$D_n = X_{1n} - X_{2n}$

N elementos

Estimar $\mu_D = \mu_1 - \mu_2$ estimador $\hat{\mu}_D = \bar{X}_1 - \bar{X}_2 = \bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}$ não enviesado

PICR

$$V(\hat{\mu}_D) = \frac{\sigma_D^2}{n} = \frac{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}{n} \qquad S_D'^2 = \frac{\sum (D_i - \overline{D})^2}{n-1}$$

$$\hat{V}(\hat{\mu}_D) = \frac{S_D'^2}{n} = \frac{S_1'^2 + S_2'^2 - 2S_{12}'}{n} \qquad S_1'^2 = \frac{\sum (X_{1i} - \overline{X}_1)^2}{n-1}$$

$$S_2'^2 = \frac{\sum (X_{2i} - \overline{X}_2)^2}{n-1}$$

$$S_{12}' = \frac{\sum (X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)}{n-1}$$

PISR

$$V(\hat{\mu}_D) = (1 - \frac{n}{N}) \frac{\sigma_D'^2}{n} = (1 - \frac{n}{N}) \times \frac{\sigma_1'^2 + \sigma_2'^2 - 2\sigma_{12}'}{n}$$

$$\hat{V}(\hat{\mu}_D) = (1 - \frac{n}{N}) \frac{S_D'^2}{n} = (1 - \frac{n}{N}) \times \frac{S_1'^2 + S_2'^2 - 2S_{12}'}{n}$$

2.4. Estimação de um rácio

Observação simultânea de duas variáveis (X,Y) com o objectivo de estimar o quociente entre essas características da população.

2.4.1. Rácio entre duas variáveis quantitativas

Estimar o quociente entre essas duas características na população, ***estimação de um rácio***

$$B = \frac{\sum_{s=1}^{N} y_s}{\sum_{s=1}^{N} x_s} = \frac{t_y}{t_x} = \frac{\mu_y}{\mu_x}$$

com base numa amostra aleatória simples de dimensão n onde se observam as duas características (X_i, Y_i) com $i = 1, 2, \ldots, n$.

EXEMPLOS:

Peso do crédito à habitação no total de crédito consumido

Peso das chamadas interurbanas no total das chamadas nacionais

Peso das despesas a com a habitação e a alimentação no rendimento disponível das famílias…

Estimador

$$\hat{B} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} = \frac{\overline{Y}}{\overline{X}}$$

Estimador enviesado para B.

No entanto para grandes amostras o enviesamento tende para zero, ou seja, $\hat{B}$ é assintoticamente não enviesado.

Cálculo do enviesamento e do EQM de $\hat{B}$

$$\hat{B}=\frac{\overline{Y}}{\overline{X}} \quad \Leftrightarrow \quad \hat{B}=B+\frac{\overline{Y}-B\overline{X}}{\overline{X}} \Leftrightarrow \hat{B}=B+\frac{\overline{Y}-B\overline{X}}{\mu_x}\times\frac{1}{\left(1+\frac{\overline{X}-\mu_x}{\mu_x}\right)}, \qquad \mu_X\neq 0.$$

Então desenvolvendo em série $(1+\delta)^{-1}=1-\delta+\delta^2-\ldots$.

$$\hat{B}=B+\frac{\overline{Y}-B\overline{X}}{\mu_x}\times\left[1-\frac{\overline{X}-\mu_x}{\mu_x}+\left(\frac{\overline{X}-\mu_x}{\mu_x}\right)^2-\ldots\right]$$

tomando a aproximação de 2ª ordem pode então dizer-se que:

$$\hat{B}-B\approx\frac{\overline{Y}-B\overline{X}}{\mu_x}\times\left[1-\frac{\overline{X}-\mu_x}{\mu_x}\right] \Rightarrow \hat{B}-B\approx\frac{\overline{Y}-B\overline{X}}{\mu_x}-\frac{\left(\overline{Y}-B\overline{X}\right)\left(\overline{X}-\mu_x\right)}{\mu_X^2}$$

$$Env(\hat{B})=E(\hat{B}-B)\approx E\left(\frac{\overline{Y}-B\overline{X}}{\mu_x}\right)-E\left(\frac{\left(\overline{Y}-B\overline{X}\right)\left(\overline{X}-\mu_x\right)}{\mu_X^2}\right)$$

Como $E\left(\frac{\overline{Y}-B\overline{X}}{\mu_x}\right)=0$ então

$$Env(\hat{B})\approx-\frac{1}{\mu_X^2}E\left(\overline{Y}\,\overline{X}-\overline{Y}\mu_X-B\overline{X}^2+B\overline{X}\mu_X\right)=$$

$$-\frac{1}{\mu_X^2}\left[E(\overline{Y}\,\overline{X})-\mu_X\mu_Y-BE(\overline{X}^2)+B\mu_X^2\right]=$$

Então, $Env(\hat{B})\approx\frac{1}{\mu_X^2}\left[B\,V(\overline{X})-Cov(\overline{Y},\overline{X})\right]$

Tomando a aproximação de 1ª ordem pode então dizer-se que $\hat{B}-B\approx\frac{\overline{Y}-B\overline{X}}{\mu_x}$

$$EQM(\hat{B})=E\left[(\hat{B}-B)^2\right]\approx E\left[\left(\frac{\overline{Y}-B\overline{X}}{\mu_x}\right)^2\right]=\frac{1}{\mu_X^2}E\left[\left(\overline{Y}-B\overline{X}\right)^2\right]$$

Enviesamento

$$Env(\hat{B})\approx\frac{1}{\mu_x^2}\left[B\times V(\overline{X})-Cov(\overline{X},\overline{Y})\right]=$$

PICR $= \frac{1}{n\mu_x^2}\left(B\times\sigma_x^2-\sigma_{xy}\right)=\frac{1}{n\mu_x^2}\left(B\times\sigma_x^2-\rho_{xy}\sigma_x\sigma_y\right)$

PISR $= (1-\frac{n}{N})\frac{\left(B\times\sigma_x'^2-\sigma_{xy}'\right)}{n\mu_x^2}=(1-\frac{n}{N})\frac{\left(B\times\sigma_x'^2-\rho_{xy}\sigma_x'\sigma_y'\right)}{n\mu_x^2}$

onde $\rho_{xy}=\frac{\sigma_{xy}}{\sigma_x\sigma_y}=\frac{\sigma_{xy}'}{\sigma_x'\sigma_y'}$ coeficiente de correlação entre *x* e *y* .

O enviesamento de $\hat{B}$ diminui quando:

- A dimensão da amostra aumenta ($n \to \infty$);
- A taxa de amostragem aumenta ($\frac{n}{N} \to 1$);
- Aumenta o coeficiente de correlação entre x e y ($\rho_{xy} \to 1$);
- A variável X tem média elevada ou desvio padrão pequeno;
- Quanto menor for o coeficiente de variação de X ($CV_x \to 0$).

O enviesamento é igual a zero, ou seja, $\hat{B}$ é não enviesado se:

$$B \times \sigma_x^2 - \sigma_{xy} = 0 \qquad \text{ou seja} \qquad \frac{\mu_y}{\mu_x} = \frac{\sigma_{xy}}{\sigma_x^2} \Leftrightarrow \frac{\mu_y}{\mu_x} = \frac{\sigma'_{xy}}{\sigma'^2_x}$$

Erro Quadrático Médio:

$$EQM(\hat{B}) = E\left[(\hat{B} - B)^2\right] \approx E\left[\frac{(\bar{Y} - B\bar{X})^2}{\mu_x^2}\right] = \frac{1}{\mu_x^2} E\left[(\bar{Y} - B\bar{X})^2\right]$$

PICR $$\approx \frac{1}{n\mu_x^2} \frac{\sum_{s=1}^{N}(y_s - Bx_s)^2}{N} = \frac{\sigma_y^2 + B^2\sigma_x^2 - 2B\sigma_{xy}}{n\mu_x^2}$$

PISR $$\approx (1 - \frac{n}{N}) \frac{1}{n\mu_x^2} \frac{\sum_{s=1}^{N}(y_s - Bx_s)^2}{N-1} = (1 - \frac{n}{N}) \frac{\sigma'^2_y + B^2\sigma'^2_x - 2B\sigma'_{xy}}{n\mu_x^2}$$

Esta aproximação para o EQM será tanto melhor quanto maior for a dimensão da amostra ($n \geq 30$ mais ou menos) e quanto menores forem os coeficientes de variação de $\bar{X}$ e $\bar{Y}$ ($CV(\bar{X}) \leq 0{,}1$ e $CV(\bar{Y}) \leq 0{,}1$). **Não se verificando estas condições** então a aproximação poderá **subestimar** fortemente o verdadeiro valor.

Estimador para o EQM: se PISR

$$E\hat{Q}M(\hat{B}) \approx (1 - \frac{n}{N}) \frac{1}{n\bar{X}^2} \sum_{i=1}^{n} \frac{(Y_i - \hat{B}X_i)^2}{n-1}$$

$$\approx (1 - \frac{n}{N}) \frac{\hat{B}^2}{n(n-1)} \sum_{i=1}^{n} \left(\frac{Y_i}{\bar{Y}} - \frac{X_i}{\bar{X}}\right)^2$$

$$\to \approx (1 - \frac{n}{N}) \frac{1}{n\bar{X}^2} \frac{\sum_{i=1}^{n} Y_i^2 + \hat{B}^2 \sum_{i=1}^{n} X_i^2 - 2\hat{B} \sum_{i=1}^{n} X_i Y_i}{n-1}$$

$$\to \approx (1 - \frac{n}{N}) \frac{1}{n\bar{X}^2} \left(S'^2_Y + \hat{B}^2 S'^2_X - 2\hat{B}\, r_{XY} S'_X S'_Y\right)$$

$$\approx (1 - \frac{n}{N}) \frac{\hat{B}^2}{n} \left(CV'^2_Y + CV'^2_X - 2\, r_{XY} CV'_X CV'_Y\right)$$

onde

$r_{XY} = \frac{S_{xy}}{S_x S_y} = \frac{S'_{xy}}{S'_x S'_y}$ coeficiente de correlação na amostra

$CV'_X = \frac{S'_X}{\overline{X}}$ e $CV'_Y = \frac{S'_Y}{\overline{Y}}$ coeficientes de variação na amostra

Distribuição assintótica de $\hat{B}$

$$\frac{\hat{B} - B}{\sqrt{E\hat{Q}M(\hat{B})}} \dot{\sim} N(0;1)$$

Assim o intervalo de confiança a *100(1-α)%* para *B* será:

$$\hat{B} \pm z_{\alpha/2}\sqrt{E\hat{Q}M(\hat{B})}$$

2.4.2. Média dos quocientes ou rácio médio

$$\mu_B = \frac{\sum_{s=1}^{N} B_s}{N} = \frac{1}{N}\sum_{s=1}^{N}\frac{y_s}{x_s}$$

com base numa amostra aleatória simples de dimensão *n* onde se observam as duas características (X_i, Y_i) com *i =1, 2, …, n.*

EXEMPLOS:

Peso médio, por agência, do crédito à habitação no total de crédito concedido;

Peso médio, por cliente, das chamadas interurbanas nas chamadas nacionais efectuadas mensalmente;

Peso médio, por família, das despesas com a habitação no rendimento disponível da família…

Estimador:

$\hat{\mu}_{\hat{B}} = \overline{B} = \sum_{i=1}^{n}\frac{B_i}{n} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{X_i}$ Estimador não enviesado para μ_B

Tendo-se:

	PICR	PISR
$V(\hat{\mu}_{\hat{B}}) =$	$\frac{\sigma_B^2}{n}$	$(1-\frac{n}{N})\frac{\sigma_B'^2}{n}$
$\hat{V}(\hat{\mu}_{\hat{B}}) =$	$\frac{S_B'^2}{n}$	$(1-\frac{n}{N})\frac{S_B'^2}{n}$

onde $S_B'^2 = \frac{\sum_{i=1}^{n}(B_i - \overline{B})^2}{n-1} = \frac{\sum_{i=1}^{n} B_i^2 - n\overline{B}^2}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\frac{Y_i}{X_i}\right)^2 - n\overline{B}^2\right]$

Utilização de $\hat{\mu}_B$ na estimação de B:

$$E(\overline{B}) - B = \mu_B - B = \frac{\sum_{s=1}^{N} B_s}{N} - \frac{\sum_{s=1}^{N} y_s}{\sum_{s=1}^{N} x_s} = \frac{1}{\sum_{s=1}^{N} x_s}(\mu_x \sum_{s=1}^{N} B_s - \sum_{s=1}^{N} y_s) \qquad (y_s = B_s x_s)$$

$$= -\frac{1}{t_x}(\sum_{s=1}^{N} B_s x_s - \mu_x \sum_{s=1}^{N} B_s) = -\frac{1}{t_x}\sum_{s=1}^{N} B_s (x_s - \mu_x) = -\frac{N-1}{t_x}\sigma'_{Bx}$$

Assim, utilizando $\overline{B}$, este é um estimador enviesado de *B*. O seu enviesamento pode ser estimado através de

	PICR	PISR
$E\hat{n}v =$	$-\frac{N}{t_x}S'_{Bx}$	$-\frac{N-1}{t_x}S'_{Bx}$

onde $S'_{Bx} = \frac{\sum_{i=1}^{n} B_i X_i - n\overline{B}\overline{X}}{n-1} = \frac{\sum_{i=1}^{n} Y_i - n\overline{B}\overline{X}}{n-1} = \frac{n}{n-1}(\overline{Y} - \overline{B}\overline{X})$

Se t_x for conhecido poder-se-á corrigir o enviesamento obtendo assim $\hat{B}_{HR}$, estimador não enviesado para *B*, conhecido por **estimador de Hartley-Ross**:

	PICR	PISR
$\hat{B}_{HR} =$	$\overline{B} + \frac{N}{t_x}\frac{n}{n-1}(\overline{Y} - \overline{B}\overline{X})$	$\overline{B} + \frac{N-1}{t_x}\frac{n}{n-1}(\overline{Y} - \overline{B}\overline{X})$

$\hat{V}(\hat{B}_{HR}) = E\hat{Q}M(\overline{B}) = \frac{S_B'^2}{n} + \frac{n^2}{\mu_X^2 (n-1)^2}(\overline{Y} - \overline{B}\overline{X})^2$ **PICR**

$= (1 - \frac{n}{N})\frac{S_B'^2}{n} + \frac{(N-1)^2 n^2}{N^2 \mu_X^2 (n-1)^2}(\overline{Y} - \overline{B}\overline{X})^2$ **PISR**

Se μ_x não for conhecido substitui-se pela média da amostra $\overline{X}$.

2.5. Características qualitativas: estimação de uma proporção

(População de Bernoulli)

Estimação do número total ou da proporção de elementos, pertencentes a determinada população, que possuem certa característica ou atributo.

$x_1 \quad x_2 \ldots \quad x_s \ldots \quad x_N$ onde $x_s = \begin{cases} 1 & \text{se } U_s \text{ tem a característica} \\ 0 & \text{se } U_s \text{ não tem a característica} \end{cases}$

Proporção $\mu = p = \dfrac{\sum_{s=1}^{N} x_s}{N}$ **estimador** $\hat{p} = \overline{X} = \sum X_i / n$

Nº total $t = \sum_{s=1}^{N} x_s$ **estimador** $\hat{T} = N\overline{X}$

- **Variâncias dos estimadores**

	PICR	*PISR*
$\sigma^2_{\hat{p}} = V(\hat{p}) =$	$\dfrac{pq}{n}$	$= \dfrac{N-n}{N-1}\dfrac{pq}{n} = (1-\dfrac{n}{N})\dfrac{Npq}{(N-1)n}$
$\sigma^2_{\hat{T}} = V(\hat{T}) =$	$\dfrac{N^2 pq}{n}$	$= \dfrac{N-n}{N-1}\dfrac{N^2 pq}{n}$

- **Estimadores para as variâncias dos estimadores**

$$S'^2 = \frac{\sum X_i^2 - n\overline{X}^2}{n-1} = \frac{n\overline{X} - n\overline{X}^2}{n-1} = \frac{n\overline{X}(1-\overline{X})}{n-1} \quad \text{neste caso } \sum X_i = \sum X_i^2$$

E assim os erros padrão dos estimadores são dados por:

	Erro Padrão (*EP*)	
	PICR	**PISR**
Proporção	$\sqrt{\dfrac{\overline{x}(1-\overline{x})}{n-1}}$	$\sqrt{1-\dfrac{n}{N}}\sqrt{\dfrac{\overline{x}(1-\overline{x})}{n-1}}$
Total	$\sqrt{\dfrac{N^2\overline{x}(1-\overline{x})}{n-1}}$	$\sqrt{(1-\dfrac{n}{N})\dfrac{N^2\overline{x}(1-\overline{x})}{n-1}}$

Construção de intervalos de confiança

(***n*** suficientemente grande)

$$\frac{T_n-\theta}{\sqrt{V(T_n)}} \dot{\sim} N(0;1) \quad \Rightarrow \quad Z_n = \frac{T_n-\theta}{\sqrt{\hat{V}(T_n)}} \dot{\sim} N(0;1)$$

Fixado o grau de confiança, 1-α, calcula-se o valor de $z_{\alpha/2}$: $P(-z_{\alpha/2} \le Z_n \le z_{\alpha/2}) = 1-\alpha$

Resolvendo a dupla desigualdade de em ordem a θ obtém-se o I.C. a $100(1-\alpha)\%$:

$$t_n - z_{\alpha/2}\sqrt{\hat{V}(T_n)} \le \theta \le t_n + z_{\alpha/2}\sqrt{\hat{V}(T_n)}$$

$$\left(t_n \pm z_{\alpha/2}\sqrt{\hat{V}(T_n)}\right) \quad ou \quad \left(t_n \pm z_{\alpha/2}EP(T_n)\right)$$

I.C. para a média de uma população (μ ou p)

Selecção com reposição

$$\left(\bar{x} \pm z_{\alpha/2} \frac{s'}{\sqrt{n}} \right) \qquad \left(\bar{x} \pm z_{\alpha/2} \sqrt{\frac{\bar{x}(1-\bar{x})}{n-1}} \right)$$

Selecção sem reposição

$$\left(\bar{x} \pm z_{\alpha/2} \frac{s'}{\sqrt{n}} \sqrt{(1-n/N)} \right) \qquad \left(\bar{x} \pm z_{\alpha/2} \sqrt{(1-n/N)} \sqrt{\frac{\bar{x}(1-\bar{x})}{n-1}} \right)$$

2.6. Dimensão da amostra

1º. Fixar o grau de confiança *(1-α)100%* nos resultados e conclusões a obter, normalmente expresso em termos de probabilidade ou de percentagem. (mais utilizados 90% ou 95%);

2º. Fixar a precisão das estimativas: erro ou desvio máximo (e) admitido (desvio ou diferença entre a estimativa e o valor real, mas desconhecido, do parâmetro)

- o parâmetro a estimar situar-se-á entre t_n-e e t_n+e (onde t_n é a estimativa obtida), ou seja, $|t_n-\theta| \leq e$.

3º. Determinar a distribuição do estimador T_n .

Na maior parte dos casos, o que se pretende estimar é a média da população e, assim, T_n será a um seu estimador não enviesado e consistente, com n suficientemente grande, poder-se-á aplicar a aproximação à distribuição normal:

$$\frac{T_n-\theta}{\sigma_{T_n}} \sim N(0;1)$$

4º. Determinar a dimensão da amostra n, de forma que:

$P(|T_n-\theta| \leq e) = 1-\alpha$, onde T_n - é a estatística, base da estimativa t_n,

$$P\left[\left(\frac{|T_n-\theta|}{\sigma_{T_n}} \leq \frac{e}{\sigma_{T_n}}\right)\right] = 1-\alpha \quad \Rightarrow \quad 2\Phi\left(e/\sigma_{T_n}\right)-1 = 1-\alpha$$

$$\Rightarrow \quad \Phi\left(e/\sigma_{T_n}\right) = 1-\frac{\alpha}{2} \quad \Rightarrow \quad \frac{e}{\sigma_{T_n}} = z_{\alpha/2}$$

$$\frac{e^2}{\sigma^2_{T_n}} = z^2_{\alpha/2}$$

onde $z_{\alpha/2} : \Phi(z_{\alpha/2}) = 1-\frac{\alpha}{2}$ $\quad (1-\alpha = 0{,}95 \Rightarrow z_{\alpha/2} = 1{,}96 \; ; \; se \; 1-\alpha = 0{,}90 \Rightarrow z_{\alpha/2} = 1{,}645)$

2.6.1. Características quantitativas

Quando o estimador é $\overline{X}$

$$\frac{e^2}{\sigma^2_{\overline{X}}} = z^2_{\alpha/2} \qquad \textbf{[1]}$$

Selecção com reposição

$$\sigma^2_{\overline{X}} = \sigma^2 / n \quad \Rightarrow \quad n = \frac{\sigma^2 \; z^2_{\alpha/2}}{e^2} \quad (= n_o)$$

Selecção sem reposição

$$\sigma^2_{\overline{X}} = \frac{N-n}{N}\sigma'^2 / n \quad \Rightarrow$$ substituindo em [1] e resolvendo em ordem a n

$$n = \frac{\sigma'^2 \; z^2_{\alpha/2}}{e^2 + \frac{\sigma'^2 \; z^2_{\alpha/2}}{N}} = \frac{N\sigma'^2 \; z^2_{\alpha/2}}{Ne^2 + \sigma'^2 \; z^2_{\alpha/2}}$$

Na prática calcula-se primeiro a dimensão da amostra como se fosse com reposição (n_o). Se n_o/N for muito pequeno (inferior a 0,05) então n_o é uma aproximação satisfatória para n.

Caso contrário calcula-se o valor de n, pois também se pode escrever:

$$n = \frac{\frac{z^2_{\alpha/2}\,\sigma^2}{e^2}}{1 + (\frac{z^2_{\alpha/2}\,\sigma^2}{e^2} - 1)/N} = \frac{n_o}{1 + \frac{n_o - 1}{N}} \approx \frac{n_o}{1 + \frac{n_o}{N}} \qquad \text{com} \qquad n_o = \frac{\sigma^2 \; z^2_{\alpha/2}}{e^2}$$

Parâmetro desconhecido (σ^2 ou σ'^2)

- Estimar usando a variância corrigida de amostras de outros estudos sobre a mesma característica ou mesmo de uma pesquisa exploratória a este estudo;
- Explicitar o erro como função da variância, por exemplo $e = 0{,}04\sigma \Rightarrow n = \frac{z^2_{\alpha/2}}{0{,}04^2}$;

Se for **fixada a precisão relativa**, podem obter-se expressões alternativas para a dimensão da amostra, pois o desvio máximo em termos absolutos (e) é igual a $d\mu$, onde d é o erro relativo, e então substituindo nas expressões anteriores obtém-se:

Selecção com reposição

$$n_o = \frac{CV^2 \; z^2_{\alpha/2}}{d^2} \qquad \text{com} \quad CV = \frac{\sigma}{\mu}$$

Selecção sem reposição

$$n = \frac{N \times CV'^2 \times z^2_{\alpha/2}}{Nd^2 + CV'^2 \times z^2_{\alpha/2}} = \frac{\frac{CV^2 \times z^2_{\alpha/2}}{d^2}}{1+(\frac{CV^2 \times z^2_{\alpha/2}}{d^2})/N} = \frac{n_o}{1+\frac{n_o-1}{N}} \approx \frac{n_o}{1+\frac{n_o}{N}} \quad \text{onde } CV' = \frac{\sigma'}{\mu}$$

2.6.2. Características qualitativas

(Variável de Bernoulli)

como $\sigma^2 = p(1-p) \le \frac{1}{4}$ então $n_o = \frac{p(1-p)z^2_{\alpha/2}}{e^2} \le \frac{z^2_{\alpha/2}}{4e^2}$ majoração

Se for **<u>fixada a precisão relativa</u>** e tratando-se de ***populações de Bernoulli***

$n_o = \frac{z^2_{\alpha/2}}{d^2}\frac{q}{p}$ já não sendo possível a majoração.

<u>EXEMPLO</u>:

Uma agência de publicidade afirma que numa campanha publicitária feita recentemente atingiu cerca de 30% das famílias de certa localidade. A empresa interessada (que pagou a campanha) duvida dessa percentagem e resolve fazer um inquérito por amostragem junto das 30000 famílias dessa localidade para verificar da autenticidade da afirmação. Qual deve ser a dimensão duma amostra aleatória simples (sem reposição) para que, com um grau de confiança de 95%, a estimativa obtida tenha um erro máximo de 5%.

Pretende-se obter $n : P(|T_n - p| < 0{,}05) = 0{,}95$ numa amostra PISR.

$$n_o = \frac{z^2_{\alpha/2}\, p(1-p)}{e^2} \qquad \text{com } p \le 0{,}3,\ z_{\alpha/2} = 1{,}96 \text{ e } e = 0{,}05$$

$$n_o = \frac{1{,}96^2\, p(1-p)}{0{,}05^2} \le \frac{1{,}96^2 \times 0{,}3 \times 0{,}7}{0{,}05^2} \approx 323$$

Taxa de amostragem de 323/30 000 ≈ 0,01 (1%).

Assim,

$$n = \frac{n_o}{1+(n_o-1)/N} = \frac{323}{1+322/30000} \approx 320 \text{ famílias} \quad (-3 \text{ do que com reposição})$$

Se na localidade houvesse 3000 famílias, a dimensão da amostra sem reposição seria de 292, contra as 323 com reposição, justificando-se a aplicação das c.p.f. .

3. AMOSTRAGEM ESTRATIFICADA

Leitura obrigatória: capítulo 3 do livro "Sampling: Design and Analysis", Sharon L. Lohr

3.1. Conceitos e notações. Estimadores e suas propriedades

3.1.1. Conceitos e notações

Dividir os N elementos da população em H grupos homogéneos, chamados estratos. Seleccionar independentemente uma amostra aleatória em cada estrato.

Estimar: média ou total da população e a variância dos seus estimadores.

ESTRATOS

	1	2	...	h	...	H	
Na população: Dimensão	N_1	N_2	...	N_h	...	N_H	$N=\sum_{h=1}^{H} N_h$
Média	μ_1, p_1	μ_2, p_2	...	μ_h, p_h	...	μ_H, p_H	$\mu=\sum_{h=1}^{H}\frac{N_h}{N}\mu_h \;;\; p=\sum_{h=1}^{H}\frac{N_h}{N}p_h$;
Total	t_1	t_2		t_h		t_H	$t_x=\sum_{h=1}^{H} t_h$
Variância	σ_1^2	σ_2^2	...	σ_h^2	...	σ_H^2	$\sigma^2=\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2+\sigma_o^2$
	$\sigma_1'^2$	$\sigma_2'^2$		$\sigma_h'^2$		$\sigma_H'^2$	$onde\ \sigma_o^2=\sum_{h=1}^{H}\frac{N_h}{N}(\mu_h-\mu)^2$
Na amostra: Dimensão	n_1	n_2	...	n_h	...	n_H	$n=\sum_{h=1}^{H} n_h$
Média, freq. obs.	$\bar{X}_1$	$\bar{X}_2$	...	$\bar{X}_h$	...	$\bar{X}_H$	$\bar{X}$
Variância	$S_1'^2$	$S_2'^2$	...	$S_h'^2$	...	$S_H'^2$	S'^2

Nos estratos:

$$\mu_h = E(X_h) = \frac{\sum_{s=1}^{N_h} X_{hs}}{N_h}; \qquad \sigma_h^2 = V(X_h) = \frac{\sum_{s=1}^{N_h}(X_{hs}-\mu_h)^2}{N_h} \qquad \text{população}$$

$$\bar{X}_h = \frac{1}{n_h}\sum_{j=1}^{n_h} X_{hj}; \qquad S_h'^2 = \frac{1}{n_h-1}\sum_{j=1}^{n_h}(X_{hj}-\bar{X}_h)^2 \qquad \text{amostra}$$

Quatro questões fundamentais a responder:

- Regras a seguir na divisão da população: ***número*** de estratos e seus ***limites***?
- Estabelecidos os estratos e fixadas a precisão e o grau de confiança para as estimativas, como determinar a ***dimensão da amostra global***?
- Determinada a dimensão da amostra global, ***dimensionar as amostras*** a recolher ***de cada estrato***?
- Definidas as dimensões das amostras a seleccionar aleatoriamente dos estratos, quais os ***estimadores a utilizar*** para os parâmetros?

3.1.2. Estimadores e suas propriedades

A. <u>Nos estratos</u>

Para a Média $\boxed{\hat{\mu}_h=\overline{X}_h=\frac{1}{n_h}\sum_{j=1}^{n_h}X_{hj}}$ para $h = 1, 2, ..., H$

Estimador não enviesado $E(\overline{X}_h)=\mu_h$

	PICR	**PISR**	
$V(\overline{X}_h)=$	$\frac{\sigma_h^2}{n_h}$	$(1-\frac{n_h}{N_h})\frac{\sigma_h'^2}{n_h}$	$h = 1, 2, ..., H$
$\hat{V}(\overline{X}_h)=$	$\frac{S_h'^2}{n_h}$	$(1-\frac{n_h}{N_h})\frac{S_h'^2}{n_h}$	$h = 1, 2, ..., H$

Para o Total $\boxed{\hat{T}_h=N_h\overline{X}_h=\frac{N_h}{n_h}\sum_{j=1}^{n_h}X_{hj}=\frac{1}{\pi_h}\sum_{j=1}^{n_h}X_{hj}}$ para $h = 1, 2, ..., H$ $(\pi_h=\frac{n_h}{N_h})$

Estimador não enviesado $E(\hat{T}_h)=t_h$

	PICR	**PISR**	
$V(\hat{T}_h)=$	$N_h^2\frac{\sigma_h^2}{n_h}$	$(1-\frac{n_h}{N_h})\frac{N_h^2\sigma_h'^2}{n_h}$	$h = 1, 2, ..., H$
$\hat{V}(\hat{T}_h)=$	$N_h^2\frac{S_h'^2}{n_h}$	$(1-\frac{n_h}{N_h})\frac{N_h^2S_h'^2}{n_h}$	$h = 1, 2, ..., H$

B. <u>Na população</u>

Para a Média

$\boxed{\hat{\mu}_E=\sum_{h=1}^{H}\frac{N_h}{N}\overline{X}_h=\sum_{h=1}^{H}W_h\overline{X}_h}$ ***média ponderada das médias dos estratos***
(ponderação igual ao peso do estrato na população)

Estimador não enviesado e consistente para μ.

$$E(\hat{\mu}_E)=\sum_{h=1}^{H}\frac{N_h}{N}E(\overline{X}_h)=\sum_{h=1}^{H}\frac{N_h}{N}\mu_h=\mu$$

com variância,

$$V(\hat{\mu}_E) = \sum_{h=1}^{H} \frac{N_h^2}{N^2} V(\overline{X}_h) = \sum_{h=1}^{H} W_h^2 \, V(\overline{X}_h) =$$

	PICR	**PISR**
$V(\hat{\mu}_E) =$	$\sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{n_h}$	$= \sum_{h=1}^{H} W_h^2 \left(1-\frac{n_h}{N_h}\right) \frac{\sigma_h'^2}{n_h}$
$\hat{V}(\hat{\mu}_E) =$	$\sum_{h=1}^{H} W_h^2 \frac{S_h'^2}{n_h}$	$= \sum_{h=1}^{H} W_h^2 \left(1-\frac{n_h}{N_h}\right) \frac{S_h'^2}{n_h}$

Em populações de Bernoulli $\Rightarrow$ $\overline{X}_h$ é a frequência relativa observada no estrato h e $\hat{\mu}$ é uma ***média ponderada das frequências relativas dos estratos***

$$\hat{p}_E = \sum_{h=1}^{H} \frac{N_h}{N} \overline{X}_h = \sum_{h=1}^{H} W_h \overline{X}_h$$

	PICR	**PISR**
$V(\hat{p}_E) =$	$\sum_{h=1}^{H} W_h^2 \frac{p_h(1-p_h)}{n_h}$	$= \sum_{h=1}^{H} W_h^2 \frac{N_h-n_h}{N_h-1} \frac{p_h(1-p_h)}{n_h}$
$\hat{V}(\hat{p}_E) =$	$\sum_{h=1}^{H} W_h^2 \frac{\overline{X}_h(1-\overline{X}_h)}{n_h-1}$	$= \sum_{h=1}^{H} W_h^2 \left(1-\frac{n_h}{N_h}\right) \frac{\overline{X}_h(1-\overline{X}_h)}{n_h-1}$

Para o Total

$$\hat{T}_E = \sum_{h=1}^{H} \hat{T}_h = \sum_{h=1}^{H} N_h \overline{X}_h = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{j=1}^{n_h} X_{hj} = \sum_{h=1}^{H} \sum_{j=1}^{n_h} \frac{1}{\pi_{hj}} X_{hj} = \sum_{h=1}^{H} \sum_{j=1}^{n_h} \omega_{hj} X_{hj}$$

onde $\omega_{hj}(= N_h/n_h)$, ***peso amostral,*** pode ser interpretado como o número de elementos da população que são representados pelo *j-ésimo* elemento da amostra do estrato h .

Estimador não enviesado e consistente para t_x

	PICR	**PISR**
$V(\hat{T}_E) = \sum_{h=1}^{H} N_h^2 V(\overline{X}_h) =$	$\sum_{h=1}^{H} N_h^2 \frac{\sigma_h^2}{n_h}$	$\sum_{h=1}^{H} N_h^2 \left(1-\frac{n_h}{N_h}\right) \frac{\sigma_h'^2}{n_h}$
$\hat{V}(\hat{T}_E) =$	$\sum_{h=1}^{H} N_h^2 \frac{S_h'^2}{n_h}$	$\sum_{h=1}^{H} N_h^2 \left(1-\frac{n_h}{N_h}\right) \frac{S_h'^2}{n_h}$

Observação: Convém não esquecer que nas estimativas da variância dos estimadores, substituiu-se em cada estrato a variância σ_h^2 (ou $\sigma_h'^2$) pela correspondente variância amostral $S_h'^2$, o que pressupõe em **cada estrato amostras de no mínimo dois elementos**.

Intervalos de confiança

Se tivermos grandes amostras em cada estrato ou se existir um grande número de estratos (mesmo com amostras de menor dimensão), uma aproximação para os IC a *100(1-α)%*:

para a média

$$(\hat{\mu}_E \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{\mu}_E)}\ )$$

para uma proporção

$$p \in (\hat{p}_E \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{p}_E)}\ )$$

para o total

$$(\hat{t}_E \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{T}_E)}\ )$$

Alguns investigadores em vez da normal utilizam uma t de Student com *(n-H)* gl.

EXEMPLO:

Estimar a proporção de fumadores numa população com base numa amostra estratificada de 1000 indivíduos. Atendendo às idades, dividiu-se a população em 6 estratos.

Estratos (Grupos etários)	Peso na população (%) $(N_h/N)\times 100$	Dimensão da subamostra (n_h)	Frequência Relativa $(\bar{x}_h)$
0 – 10	20	200	0,00
11 – 15	12	120	0,05
16 – 20	13	130	0,25
21 – 50	40	400	0,40
51 – 70	10	100	0,25
mais de 70	5	50	0,15

O intervalo de confiança a 95% para a proporção de fumadores em toda a população será:

$$p \in \left(\hat{p}_E - z_{\alpha/2}\sqrt{\hat{V}(\hat{p})}\ ,\ \hat{p}_E + z_{\alpha/2}\sqrt{\hat{V}(\hat{p}_E)} \right)$$

$$\hat{p}_E = \sum_{h=1}^{6} \frac{N_h}{N}\bar{x}_h = 0{,}20\times 0 + 0{,}12\times 0{,}05 + 0{,}13\times 0{,}25 + 0{,}4\times 0{,}4 + 0{,}1\times 0{,}25 + 0{,}05\times 0{,}15 = 0{,}231$$

$$\hat{V}(\hat{p}_E) = \sum_{h=1}^{6} \frac{N_h^2}{N^2}\,\frac{\bar{x}_h(1-\bar{x}_h)}{n_h - 1} = 0 + 0{,}12^2\,\frac{0{,}05\times 0{,}95}{119} + 0{,}13^2\,\frac{0{,}25\times 0{,}75}{129} +$$

$$+ 0{,}4^2\,\frac{0{,}4\times 0{,}6}{399} + 0{,}1^2\,\frac{0{,}25\times 0{,}75}{99} + 0{,}05^2\,\frac{0{,}15\times 0{,}85}{49} =$$

$$= \mathbf{0{,}000152}$$

I.C. a *95%* para *p* será:

$$(0{,}231 \pm 1{,}96\sqrt{0{,}000152}\ ) \quad \Rightarrow \quad \mathbf{(0{,}207\ ;\ 0{,}255)}$$

pelo que podemos afirmar, com uma confiança de 95%, que a percentagem de fumadores dessa população se situa entre os 20,7 e os 25,5% .

3.2. Quantificação da amostra e eficácia da estratificação

3.2.1. Quantificação das amostras dos estratos

Afixação proporcional

dividir n proporcionalmente ao número de unidades estatísticas de cada estrato. Deverá verificar-se, para cada um dos H estratos:

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots = \frac{n_H}{N_H} = \frac{n}{N} \qquad com \quad n_1 + n_2 + \ldots + n_H = n \qquad e \quad N_1 + N_2 + \ldots + N_H = N$$

Obtém-se,

$$\boxed{n_h = n\frac{N_h}{N}} \quad \text{para} \quad h = 1, 2, \ldots, H$$

Afixação óptima

os n_h determinam-se de forma a que seja mínima a variância do estimador da média da população, utilizado neste tipo de amostragem.

Afixação de Neyman

Determinar os n_h: Minimizar $V(\hat{\mu}_E) = \sum_{h=1}^{H} \frac{N_h^2}{N^2}\frac{\sigma_h^2}{n_h}$ ou $= \sum_{h=1}^{H} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{\sigma_h'^2}{n_h}$

sujeito a $n_1 + n_2 + \ldots + n_H = n$ (n fixo)

Resolvendo este problema de minimização com uma restrição obtém-se:

$$\boxed{n_h = n\frac{N_h\sigma_h}{\sum_{h=1}^{H} N_h\sigma_h}} \quad \text{para} \quad h = 1, 2, \ldots, H$$

Com restrição orçamental

Determinar os n_h: Minimizar $V(\hat{\mu}_E) = \sum_{h=1}^{H} \frac{N_h^2}{N^2}\frac{\sigma_h^2}{n_h}$ ou $= \sum_{h=1}^{H} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{\sigma_h'^2}{n_h}$

sujeito a $C_o + C_1 n_1 + C_2 n_2 + \ldots + C_H n_H = C$

Procedendo como acima obtém-se:

$$\boxed{n_h = \frac{N_h\sigma_h/\sqrt{C_h}}{\sum_{h=1}^{H} N_h\sigma_h/\sqrt{C_h}}\, n = \frac{N_h\sigma_h/\sqrt{C_h}}{\sum_{h=1}^{H} N_h\sigma_h\sqrt{C_h}}\,(C - C_0)} \quad h = 1, 2, \ldots, H$$

A afixação óptima envolve duas dificuldades:

- normalmente **não são conhecidos os desvios padrões dos vários estratos**;
- na maior parte dos casos, pretendem-se estimar por amostragem os parâmetros da distribuição de **várias variáveis simultaneamente**, acontecendo que a afectação óptima para o parâmetro de uma variável geralmente diferirá da óptima para outro.

Muitas vezes, a estratificação é feita com base em variáveis caracterizadoras da dimensão das unidades estatísticas (nº de trabalhadores, volume de vendas, nº de adultos na família, nº de habitantes,...), pois as características quantitativas a estudar estão, normalmente altamente correlacionadas com a dimensão das unidades. Assim d*etermina-se a afixação óptima para a estimação da dimensão média das unidades*, esperando que em virtude da alta correlação existente entre a dimensão e as características a estudar, essa repartição também seja boa para a estimação das suas médias.

A imputação da dimensão aos estratos é feita, muitas vezes, de acordo com a seguinte regra prática: **a amostra é repartida entre os estratos proporcionalmente à soma da variável utilizada para a estratificação**.

EXEMPLO:

Queremos realizar um inquérito por amostragem a 1000 estabelecimentos industriais, para obter informações sobre o montante investido em inovação tecnológica. O universo foi dividido em dois estratos:

Estrato 1 - estabelecimentos com menos de 50 trabalhadores
Estrato 2 - estabelecimentos com 50 ou mais trabalhadores

Para o conjunto dos estabelecimentos conhecem-se os seguintes dados:

Estrato	Nº de Estabelecimentos	Nº trabalhadores	Dimensão da amostra AP	APE
1	10 460 (0,799)	210 834 (0,268)	799	268
2	2 638 (0,201)	575 772 (0,732)	201	732
Total	13 098	786 606	1000	1000

Com os dados disponíveis não é possível utilizar a afixação óptima, portanto ou determinamos a repartição dos 1000 estabelecimentos da amostra pelos estratos por afixação proporcional (AP) ou então proporcionalmente à soma da variável que serviu de base à estratificação (APE), e que foi o número de trabalhadores, uma das variáveis que caracterizam a dimensão dos estabelecimentos.

3.2.2. Variância dos estimadores na afixação proporcional e na óptima

Estimador para a média de uma característica quantitativa

	PICR	PISR	
$V(\hat{\mu}_E) =$	$\sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{\sigma_h^2}{n_h}$	$= \sum_{h=1}^{H} \frac{N_h^2}{N^2} (1-\frac{n_h}{N_h}) \frac{\sigma_h'^2}{n_h}$	**[1]**

Com afixação proporcional, $n_h = n\frac{N_h}{N}$ e substituindo em **[1]**

	PICR	PISR
$V(\hat{\mu}_{EP}) =$	$\frac{1}{n}\sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2$	$= \frac{1}{n}(1-\frac{n}{N})\sum_{h=1}^{H} \frac{N_h}{N} \sigma_h'^2$

Com afixação óptima, $n_h = n\frac{N_h\sigma_h}{\sum_{h=1}^{H} N_h\sigma_h}$ substituindo em **[1]**

	PICR	PISR
$V(\hat{\mu}_{EO}) =$	$\frac{1}{n}\left(\sum_{h=1}^{H} \frac{N_h}{N} \sigma_h\right)^2$	$= \frac{1}{n}\left(\sum_{h=1}^{H} \frac{N_h}{N} \sigma_h'\right)^2 - \frac{1}{N}\sum_{h=1}^{H} \frac{N_h}{N} \sigma_h'^2$

Caso de uma proporção

	PICR	PISR	
$V(\hat{p}_E) =$	$\sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{p_h(1-p_h)}{n_h}$	$\sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{N_h-n_h}{N_h-1} \frac{p_h(1-p_h)}{n_h}$	**[1]**

Com afixação proporcional, $n_h = n\frac{N_h}{N}$ e substituindo **[1]**

$V(\hat{\mu}_{EP}) = \frac{1}{n}\sum_{h=1}^{H} \frac{N_h}{N} p_h(1-p_h)$ $\quad \frac{1}{n}(1-\frac{n}{N})\sum_{h=1}^{H} \frac{N_h}{N}\frac{N_h}{N_h-1} p_h(1-p_h)$ como $\frac{N_h}{N_h-1} \approx 1$

$\approx \frac{1}{n}(1-\frac{n}{N})\sum_{h=1}^{H} \frac{N_h}{N} p_h(1-p_h)$

Com afixação óptima, $n_h = n\frac{N_h\sigma_h}{\sum_{h=1}^{H} N_h\sigma_h}$ substituindo em **[1]**

PICR	PISR
$V(\hat{\mu}_{EO}) = \frac{1}{n}\left(\sum_{h=1}^{H} \frac{N_h}{N} \sqrt{p_h(1-p_h)}\right)^2$	$V(\hat{\mu}_{EO}) \approx \frac{1}{n}\left(\sum_{h=1}^{H} \frac{N_h}{N} \sqrt{p_h(1-p_h)}\right)^2 - \frac{1}{N}\sum_{h=1}^{H} \frac{N_h}{N} p_h(1-p_h)$

3.2.3. Dimensão global da amostra

$n : P(|T_n - \theta| < e) = 1 - \alpha$, com n grande $\dfrac{T_n - \theta}{\sigma_{T_n}} \dot{\sim} N(0;1)$

Então,

$$\frac{e^2}{\sigma^2_{T_n}} = z^2_{\alpha/2}$$

Substituindo $\sigma^2_{T_n}$ pela sua expressão e resolvendo em ordem a n obtém-se a expressão pretendida, para o cálculo da dimensão:

<u>Caso da média de uma característica quantitativa</u>

a) Afixação proporcional

$$n = \frac{z^2_{\alpha/2} \sum_{h=1}^{H} W_h \sigma_h^2}{e^2}$$ com reposição

$$n = \frac{z^2_{\alpha/2} \sum_{h=1}^{H} W_h \sigma_h'^2}{e^2 + \frac{z^2_{\alpha/2}}{N} \sum_{h=1}^{H} W_h \sigma_h'^2}$$ sem reposição onde $W_h = \frac{N_h}{N}$

b) Afixação óptima

$$n = \frac{z^2_{\alpha/2} \left(\sum_{h=1}^{H} W_h \sigma_h \right)^2}{e^2}$$ com reposição

$$n = \frac{z^2_{\alpha/2} \left(\sum_{h=1}^{H} W_h \sigma_h' \right)^2}{e^2 + \frac{z^2_{\alpha/2}}{N} \sum_{h=1}^{H} W_h \sigma_h'^2}$$ sem reposição

Caso de uma proporção

a) Afixação proporcional

$$n = \frac{z^2_{\alpha/2}\sum_{h=1}^{H} W_h p_h(1-p_h)}{e^2} \leq \frac{z^2_{\alpha/2}}{4e^2} \qquad \text{com reposição} \qquad \text{onde } W_h = \frac{N_h}{N}$$

$$n = \frac{z^2_{\alpha/2}\sum_{h=1}^{H} W_h p_h(1-p_h)}{e^2 + \frac{z^2_{\alpha/2}}{N}\sum_{h=1}^{H} W_h p_h(1-p_h)} \qquad \text{sem reposição}$$

b) Afixação óptima

$$n = \frac{z^2_{\alpha/2}\left(\sum_{h=1}^{H} W_h \sqrt{p_h(1-p_h)}\right)^2}{e^2} \qquad \text{com reposição}$$

$$n = \frac{z^2_{\alpha/2}\left(\sum_{h=1}^{H} W_h \sqrt{p_h(1-p_h)}\right)^2}{e^2 + \frac{z^2_{\alpha/2}}{N}\sum_{h=1}^{H} W_h p_h(1-p_h)} \qquad \text{sem reposição}$$

3.2.4. Eficácia da estratificação

Vão comparar-se as variâncias calculadas anteriormente

Quando a selecção é com reposição tem-se que:

$$V(\hat{\mu}_{AS}) = \frac{\sigma^2}{n} = \frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2 + \frac{\sigma_o^2}{n} = V(\hat{\mu}_{EP}) + \frac{\sigma_o^2}{n} \quad \Rightarrow \quad V(\hat{\mu}_{EP}) \leq V(\hat{\mu}_{AS})$$

$$V(\hat{\mu}_{EO}) = \frac{1}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h\right)^2 = \frac{1}{n}\left(\sum_{h=1}^{H}\sqrt{\frac{N_h}{N}}\sqrt{\frac{N_h}{N}}\sigma_h\right)^2 \leq \frac{1}{n}\sum_{h=1}^{H}\left(\sqrt{\frac{N_h}{N}}\right)^2 \times \sum_{h=1}^{H}\left(\sqrt{\frac{N_h}{N}}\sigma_h\right)^2$$

$$\Rightarrow V(\hat{\mu}_{EO}) \leq \frac{1}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}\right)\left(\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2\right) = V(\hat{\mu}_{EP})$$

$$V(\hat{\mu}_{EO}) \leq V(\hat{\mu}_{EP}) \leq V(\hat{\mu}_{AS})$$

Quando a selecção é sem reposição tem-se que:

$$V(\hat{\mu}_{AS}) = (1-\frac{n}{N})\frac{\sigma'^2}{n} = (1-\frac{n}{N})\frac{1}{n}\left[\frac{\sum_{h=1}^{H}(N_h-1)\sigma_h'^2}{N-1} + \frac{\sum_{h=1}^{H}N_h(\mu_h-\mu)^2}{N-1}\right]$$

$$V(\hat{\mu}_{EP}) = \frac{1}{n}(1-\frac{n}{N})\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h'^2 = \frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h'^2 - \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h'^2$$

$$V(\hat{\mu}_{EO}) = \frac{1}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h'\right)^2 - \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h'^2$$

Aleatória simples *vs* estratificada com afixação proporcional

$$V(\hat{\mu}_{AS}) - V(\hat{\mu}_{EP}) = (1-\frac{n}{N})\frac{1}{n}\left[\frac{\sum_{h=1}^{H}(N_h-1)\sigma_h'^2}{N-1} + \frac{\sum_{h=1}^{H}N_h(\mu_h-\mu)^2}{N-1} - \sum_{h=1}^{H}\frac{N_h}{N}\sigma_h'^2\right]$$

Desde que $\frac{N_h-1}{N-1} \approx \frac{N_h}{N}$

$$V(\hat{\mu}_{AS}) - V(\hat{\mu}_{EP}) \approx (1-\frac{n}{N})\frac{1}{n}\frac{\sum_{h=1}^{H}N_h(\mu_h-\mu)^2}{N-1} \geq 0$$

Estratificada: afixação proporcional *vs* afixação óptima

$$V(\hat{\mu}_{EP}) - V(\hat{\mu}_{EO}) = \frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h'^2 - \frac{1}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h'\right)^2 \geq 0 \Rightarrow V(\hat{\mu}_{EP}) \geq V(\hat{\mu}_{EO})$$

pois $\left(\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h'\right)^2 = \left(\sum_{h=1}^{H}\sqrt{\frac{N_h}{N}}\sqrt{\frac{N_h}{N}}\sigma_h'\right)^2 \leq \sum_{h=1}^{H}\left(\sqrt{\frac{N_h}{N}}\right)^2 \times \sum_{h=1}^{H}\left(\sqrt{\frac{N_h}{N}}\sigma_h'\right)^2$

3.3. Efeitos de erros na grandeza dos estratos

Pode acontecer trabalhar-se com dados desactualizados o que conduz nomeadamente a valores para a dimensão dos estratos que não são os verdadeiros N_h e, assim, os pesos que se utilizam não são mais do que estimativas ($W_h' = \frac{N_h'}{N}$) para os verdadeiros pesos.

Neste caso $\hat{\mu}_E = \sum_{h=1}^{H}W_h'\overline{X}_h$ e $E(\hat{\mu}_E) = \sum_{h=1}^{H}W_h'\mu_h \neq \mu = \sum_{h=1}^{H}W_h\mu_h$, ou seja, obtém-se *uma estimativa enviesada*, sendo o seu enviesamento dado por:

$$Env(\hat{\mu}_E) = \sum_{h=1}^{H} W_h' \mu_h - \sum_{h=1}^{H} W_h \mu_h = \sum_{h=1}^{H} (W_h' - W_h)\mu_h .$$

A precisão da estimativa é medida pelo $EQM(\hat{\mu}_E) = E(\hat{\mu}_E - \mu)^2 = V(\hat{\mu}_E) + \left[\sum_{h=1}^{H} (W_h' - W_h)\mu_h\right]^2$

Quando a dimensão da amostra aumenta a Variância diminui mas o enviesamento permanece constante e, assim obtém-se uma estimativa que é menos precisa, perdendo-se o benefício da estratificação.

3.4. Construção dos estratos

Questões a responder:

(1) Como escolher a característica para organizar os estratos, isto é, como escolher a variável de estratificação?

(2) Como se determinam os limites dos estratos?

(3) Qual o número de estratos?

(1) Para uma característica ser utilizada como variável de estratificação deve:

- ter um valor conhecido para cada elemento da população,
- estar correlacionada com a variável de estudo

(2)

Regras de um modo geral dificilmente aplicáveis.

Por exemplo as estabelecidas por Dalenius (1957) que estabelece como limites óptimos dos estratos:

$\Rightarrow$ Na afixação proporcional: $y_h = \frac{\mu_h + \mu_{h+1}}{2}$ para $h = 1, 2, \ldots, H-1$

$\Rightarrow$ Na afixação óptima: $y_h : \frac{(y_h - \mu_h)^2 + \sigma_h^2}{\sigma_h} = \frac{(y_{h+1} - \mu_{h+1})^2 + \sigma_{h+1}^2}{\sigma_{h+1}}$

Vários autores desenvolveram soluções aproximadas, como por exemplo:

- $N_h \sigma_h \approx$ iguais para todos os estratos
- $N_h \mu_h \approx$ iguais (estratos com total igual)
- $(y_{h+1} - y_h) \approx$ iguais (estratos com igual amplitude)
- $W_h (y_{h+1} - y_h) \approx$ iguais
- Regra empírica e simples (muito utilizada) e desenvolvida por Dalenius e Hodges, 1959.

Conhece-se uma grelha fina de divisão da população em classes

Classes	Frequência abs (F_l)	$\sqrt{F_l}$	Acumulado	
C_1	F_1	$\sqrt{F_1}$	$\sqrt{F_1}$	
C_2	F_2	$\sqrt{F_2}$	$\sqrt{F_1}+\sqrt{F_2}$	
…	…	…	…	$\rightarrow$T/H
c_l	F_l	$\sqrt{F_l}$	$\sum_{i=1}^{l}\sqrt{F_i}$	
…	…	…	…	$\rightarrow$2T/H
c_z	F_z	$\sqrt{F_z}$	$T=\sum_{i=1}^{z}\sqrt{F_i}$	...

Para obter H estratos divido T por H e vou escolhendo como limites dos estratos os limites das classes que correspondem aos acumulados: $\frac{T}{H}$, $\frac{2T}{H}$, …, $\frac{(H-1)T}{H}$.

(3) Qual o número de estratos?

A priori se a estratificação traz ganhos de precisão existe uma tentação para multiplicar o número de estratos, mas é necessário sempre ter em conta a dimensão da amostra. Para decidir é conveniente ter em conta o seguinte:

- $\Rightarrow$ Qual a *diminuição da variância* resultante do acréscimo do número de estratos?
- $\Rightarrow$ Qual o *aumento de custos* e *complexidade* dos cálculos resultantes desse acréscimo?
- $\Rightarrow$ A dimensão das amostras nos estratos deve ser sempre suficiente para permitir estimar a variância convenientemente.

CONCLUSÃO:

- $\Rightarrow$ Não há fórmula mágica, para cálculo do número óptimo. Bom senso e experiência.
- $\Rightarrow$ Para cada variável de estratificação um número de estratos entre 2 e 10.
- $\Rightarrow$ Pode usar-se uma regra empírica para o número de classes $H=1+\frac{10}{3}log(n)$.

EXEMPLO:
Os dados do quadro seguinte dão-nos a distribuição de frequência da percentagem de empréstimos bancários para fins industriais numa população de 13435 bancos de certa região dos EUA. Como se vê a distribuição é fortemente assimétrica.

(Vamos considerar H=5 estratos)

Classes	Frequência abs (F_l)	$\sqrt{F_l}$	Acumulados	
0-5	3464	58,86	58,86	T/H = 389,46/5
				← **=77,89**
5-10	2516	50,16	109,02	
10-15	2157	46,44	155,46	
				← **155,78**
15-20	1581	39,76	195,22	
20-25	1142	33,79	229,02	
				← **233,67**
25-30	746	27,31	256,33	
30-35	512	22,63	278,96	
35-40	376	19,39	298,35	
				← **311,56**
40-45	265	16,28	314,63	
45-50	207	14,39	329,02	
50-55	126	11,22	340,24	
55-60	107	10,34	350,58	***E1 0% - 5%***
60-65	82	9,06	359,64	***E2 5% - 15%***
65-70	50	7,07	366,71	***E3 15% - 25%***
70-75	39	6,24	372,96	***E4 25% - 45%***
75-80	25	5,00	377,96	***E5 45% - 100%***
80-85	16	4,00	381,96	
85-90	19	4,36	386,32	
90-95	2	1,41	387,73	
95-100	3	1,73	389,46	← 389,46
	13435	389,46 = T		

4. UTILIZAÇÃO DE INFORMAÇÃO SUPLEMENTAR

Leitura obrigatória: capítulo 4 do livro "Sampling: Design and Analysis", Sharon L. Lohr

4.1. Estratificação a posteriori

1) Pretende-se estudar uma característica *Y* sobre uma população, com base numa amostra aleatória simples.
2) Conhece-se a distribuição de uma outra característica *X* (que se supõe muito relacionada com a primeira) para a mesma população, mas não foi possível estratificar a população segundo esta característica. No entanto a cada uma das unidades estatísticas que integraram a amostra foi inquirido o valor das duas características.

Procedimento:

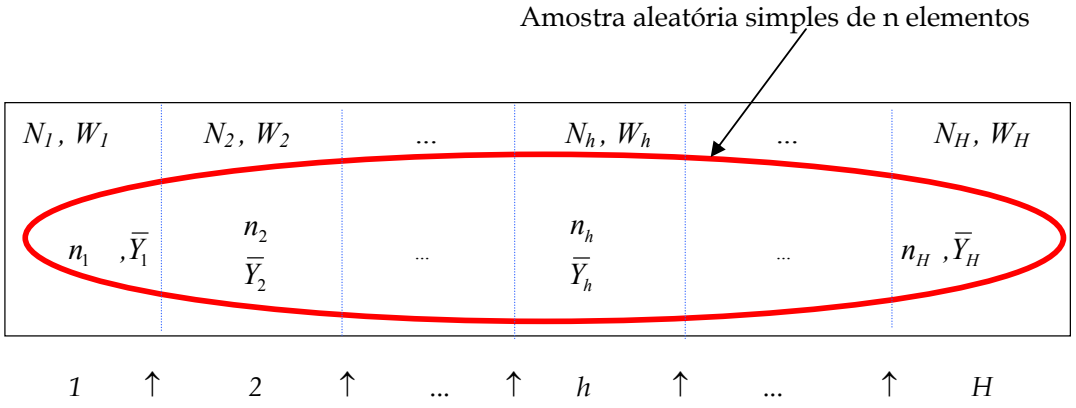
→ ***a posteriori* são definidos estratos** segundo a variável *X* e as unidades amostradas são reagrupadas segundo os valores observados de *X*.

→ ***reponderam-se* os valores amostrais de Y**, de acordo com os pesos dos estratos definidos pela distribuição de X conhecida para a população.

Exemplo:
Pretende-se estimar o gasto mensal médio em alimentação, por agregado familiar, duma determinada zona. Com base nos dados do Censo, sabe-se que a distribuição dos agregados dessa zona, consoante a dimensão do agregado, é a seguinte:

Nº de pessoas	1	2	3	4	5 ou +
% de agregados	25,7	31,2	17,5	15,6	10,0

No entanto não é possível conhecer *a priori* qual a dimensão de cada um dos agregados que pertencem à zona e, assim, não se pode utilizar tal característica como variável de estratificação.
Então como proceder?



Fronteiras traçadas a posteriori

Conhecem-se: $W_1, W_2, ..., W_h, ..., W_H$ - pesos dos estratos

O **estimador por pós-estratificação** para a média da amostra:

$$\hat{\mu}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{Y}_h = \sum_{h=1}^{H} W_h \bar{Y}_h$$

A questão tal como na estimação em subpopulações é que tanto $\bar{Y}_h$ como n_h são variáveis aleatórias.

Prova-se, no entanto, que este estimador é **não enviesado**

$$E(\hat{\mu}_{post}) = \underset{n_h}{E}\left[\underbrace{E(\hat{\mu}_{post} \mid n_h)}_{=\mu} \right] = \mu$$

Sendo a sua **variância**,

$$V(\hat{\mu}_{post}) = \underbrace{\underset{n_h}{V}\left[E(\hat{\mu}_{post} \mid n_h)\right]}_{=0} + \underset{n_h}{E}\left[V(\hat{\mu}_{post} \mid n_h)\right]$$

$$V(\hat{\mu}_{post}) \approx (1-\frac{n}{N})\frac{1}{n}\sum_{h=1}^{H} W_h \sigma_h'^2 + (1-\frac{n}{N})\frac{1}{n^2}\sum_{h=1}^{H}(1-W_h)\sigma_h'^2$$

$V(\hat{\mu}_{EP})$ + acréscimo por não ter estratificado *a priori*

podendo então afirmar-se que $V(\hat{\mu}_{post}) \geq V(\hat{\mu}_{EP})$.

Estimador para a variância:

$$\hat{V}(\hat{\mu}_{post}) \approx (1-\frac{n}{N})\frac{1}{n}\sum_{h=1}^{H} W_h S_h'^2 + (1-\frac{n}{N})\frac{1}{n^2}\sum_{h=1}^{H}(1-W_h)S_h'^2$$

Quando os n_h são razoavelmente grandes (≥30) e a dimensão da amostra global é elevada pode utilizar-se a variância da amostragem estratificada com afixação proporcional como aproximação, ou seja,

$$\hat{V}(\hat{\mu}_{post}) \approx (1-\frac{n}{N})\frac{1}{n}\sum_{h=1}^{H} W_h S_h'^2$$

- ***Comparação com a amostragem aleatória simples***

$$V(\hat{\mu}_{AS}) - V(\hat{\mu}_{post}) \approx (1-\frac{n}{N})\left[\sum_{h=1}^{H} W_h(\mu_h - \mu)^2 - \frac{1}{n}\sum_{h=1}^{H}(1-W_h)\sigma_h'^2\right]$$

Estratificação a posteriori será tanto mais justificável quanto maior for esta quantidade.

Então para utilizar a pós-estratificação:

- Variável estudada muito correlacionada com a variável de estratificação;
- Grandes amostras (*n* elevado), não é aconselhável pós-estratificar pequenas amostras;
- *(1-W_h)* devem ser pequenos, ou seja, é indesejável ter estratos a posteriori de pequena dimensão (isto é, devem-se definir poucos estratos);

- Conhecer com precisão os pesos dos estratos (W_h) que servem de base à construção da estimativa, pois o estimador só é não enviesado se os pesos forem os verdadeiros. Os problemas de enviesamento ⇒ erros padrão que não diminuem com a dimensão da amostra (estimadores que não são consistentes).

- ***Situações onde se aplica a estratificação a posteriori***

1. Existência de critérios úteis para a interpretação dos resultados (variáveis de controlo), com distribuição conhecida para a população, mas não estando a base de sondagem identificada por esses critérios não é possível organizar um plano de amostragem estratificado a priori;
2. Situações de taxas de não resposta significativas nomeadamente em certos segmentos da população; tentativa de minorar os enviesamentos provocados por distorções amostrais.

Nos questionários introduzem-se questões que servem para *descrever* a unidade estatística inquirida quanto a *certas características* que não aquela em estudo, variáveis de controlo (por exemplo: idade, estado civil, habilitações, …, nº de pessoas do agregado familiar, nº de crianças no agregado familiar, características da habitação, …)

4.2. Estimação pelo quociente

Observação simultânea de duas variáveis (X,Y) com o objectivo de estimar os parâmetros de uma das características utilizando a informação recolhida referente às duas variáveis, ***estimador por índice*** ou ***pelo quociente***.

Construção de estimadores que serão tanto mais eficazes quanto maior a correlação entre as variáveis.

Estimar a média μ_y ou o total t_y da característica Y, utilizando uma variável auxiliar X conhecida para a população (conhece-se μ_x ou t_x).

Para os ***n*** elementos que compõem a amostra aleatória simples observam-se os valores X_i e Y_i, com $i=1, 2, \ldots, n$..

O estimador por índice ou estimador quociente:

Para a média de Y, ($\hat{\mu}_{y_r}$):

$$\hat{\mu}_{y_r} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}\mu_x = \frac{\overline{Y}}{\overline{X}}\mu_x$$

Para o total de Y, ($\hat{T}_{y_r}$):

$$\hat{T}_{y_r} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}t_x = \frac{\overline{Y}}{\overline{X}}t_x$$

Enviesamento:

	PICR		PISR
$Env(\hat{\mu}_{y_r}) \approx$	$\frac{\left(B\times\sigma_x^{2}-\sigma_{xy}\right)}{n\mu_x}$	ou	$(1-\frac{n}{N})\frac{\left(B\times\sigma_x'^{2}-\sigma_{xy}'\right)}{n\mu_x}$
$Env(\hat{T}_{y_r}) \approx$	$\frac{t_x\left(B\times\sigma_x^{2}-\sigma_{xy}\right)}{n\mu_x^{2}}$	ou	$(1-\frac{n}{N})\frac{t_x\left(B\times\sigma_x'^{2}-\sigma_{xy}'\right)}{n\mu_x^{2}}$

Observações:

- Estimadores geralmente enviesados mas assintoticamente não enviesados;
- Enviesamento é nulo se e só se $\frac{\mu_y}{\mu_x} = \frac{\sigma_{xy}}{\sigma_x^2} \Leftrightarrow \mu_y = \frac{\sigma_{xy}}{\sigma_x^2}\mu_x$ o que acontece quando a regressão de *Y* sobre *X* é uma recta que passa pela origem;
- Não interessa (é mesmo prejudicial) a existência de uma correlação negativa entre *X* e *Y*, pois aumenta o enviesamento e diminuiu a precisão.

Erro Quadrático Médio:

$$EQM(\hat{\mu}_{y_r}) = \mu_x^2 EQM(\hat{B}) \qquad \text{e} \qquad EQM(\hat{T}_{y_r}) = t_x^2 EQM(\hat{B})$$

Estimador para o EQM – Precisão dos estimadores

$\hat{EQM}(\hat{\mu}_{y_r}) = \mu_x^2 \hat{EQM}(\hat{B})$ se PISR

$$\approx (1-\frac{n}{N})\frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \hat{B}X_i)^2}{n-1} = (1-\frac{n}{N})\frac{S_e'^2}{n}$$ onde $e_i = Y_i - \hat{B}X_i$ é o i-ésimo resíduo resultante do ajustamento de $y = Bx + \varepsilon$

$$\rightarrow \quad \approx (1-\frac{n}{N})\frac{\sum_{i=1}^{n}Y_i^2 + \hat{B}^2\sum_{i=1}^{n}X_i^2 - 2\hat{B}\sum_{i=1}^{n}X_iY_i}{n(n-1)}$$

$$\rightarrow \quad \approx (1-\frac{n}{N})\frac{1}{n}\left(S_Y'^2 + \hat{B}^2 S_X'^2 - 2\hat{B}\, r_{XY} S_X' S_Y'\right)$$

$$\approx (1-\frac{n}{N})\frac{\overline{Y}^2}{n}\left(CV_Y'^2 + CV_X'^2 - 2\, r_{XY} CV_X' CV_Y'\right)$$

$\hat{EQM}(\hat{T}_{y_r}) = t_x^2 \hat{EQM}(\hat{B})$

$$\rightarrow \quad \approx (1-\frac{n}{N})\frac{t_x^2}{n\overline{X}^2}\frac{\sum_{i=1}^{n}Y_i^2 + \hat{B}^2\sum_{i=1}^{n}X_i^2 - 2\hat{B}\sum_{i=1}^{n}X_iY_i}{n-1}$$ substituindo $\frac{t_X^2}{\overline{X}^2}$ por N^2

$$\approx (1-\frac{n}{N})\frac{t_x^2}{n\overline{X}^2}\left(S_Y'^2 + \hat{B}^2 S_X'^2 - 2\hat{B}\, r_{XY} S_X' S_Y'\right)$$ se *N* conhecido

Comparação da precisão de $\hat{\mu}_{y_r}$ e de $\overline{Y}$ como estimadores de μ_y

$$EQM(\overline{Y}) = V(\overline{Y}) = \frac{N-n}{Nn}\sigma_y'^2 \text{ e}$$

$$EQM(\hat{\mu}_{y_r}) = \frac{N-n}{Nn}\left(\sigma_y'^2 + B^2\sigma_x'^2 - 2B\,\rho_{xy}\sigma_x'\sigma_y'\right)$$

então, $EQM(\overline{Y}) - EQM(\hat{\mu}_{y_r}) = \frac{N-n}{Nn}B\left(2\,\rho_{xy}\sigma_x'\sigma_y' - B\sigma_x'^2\right)$

e para que $\hat{\mu}_{y_r}$ tenha maior precisão que $\overline{Y}$ é necessário que a diferença seja > 0, ou seja,

$$EQM(\overline{Y}) - EQM(\hat{\mu}_{y_r}) > 0 \quad \Leftrightarrow \quad \frac{N-n}{Nn} B\left(2\,\rho_{xy}\sigma'_x\sigma'_y - B\sigma'^2_x\right) > 0 \Leftrightarrow \quad 2\,\rho_{xy}\sigma'_x\sigma'_y > B\sigma'^2_x \quad \Leftrightarrow$$

$$\rho_{xy} > \frac{B\sigma'_x}{2\sigma'_y} \quad \Leftrightarrow \quad \boxed{\rho_{xy} > \frac{CV'_x}{2CV'_y}}$$

Se os coeficientes de variação das duas variáveis são aproximadamente iguais então $\hat{\mu}_{y_r}$ ($\hat{T}_{y_r}$) terá maior precisão que $\overline{Y}$ ($\hat{T}_y$) se $\rho_{xy} > 0{,}5$.

Quando CV_x é superior ao dobro do CV_y então $\overline{Y}$ ($\hat{T}_y$) tem sempre maior precisão que $\hat{\mu}_{y_r}$ ($\hat{T}_{y_r}$) (pois como se sabe o coeficiente de correlação nunca pode ser maior que 1).

Em conclusão para utilizar o estimador por índice:

- Deve ser possível observar simultaneamente as duas variáveis *X* e *Y*, conhecendo-se com exactidão a média ou o total da variável auxiliar.
- O coeficiente de variação de *X* não deve ser muito superior ao de *Y*.
- As variáveis *X* e *Y* devem estar positivamente correlacionadas (e quanto maior melhor)
- Devem tratar-se de grandes amostras, de modo a que os coeficientes de variação de $\overline{X}$ e $\overline{Y}$ sejam pequenos ($CV(\overline{X}) \leq 0{,}1$ e $CV(\overline{Y}) \leq 0{,}1$).

EXEMPLO: (Fonte: Tópicos de Sondagens – Paulo Gomes)

População em 1970 t_x = 22 919 milhares

População em 1980 t_y = 29351 milhares (*parâmetro a estimar*)

N=196 cidades existentes → amostra aleatória simples de *n=49* cidades

Seleccionaram-se 200 amostras aleatórias simples de *49* cidades, registando-se para cada amostra a população em 1970 e 1980 dessas 49 cidades:

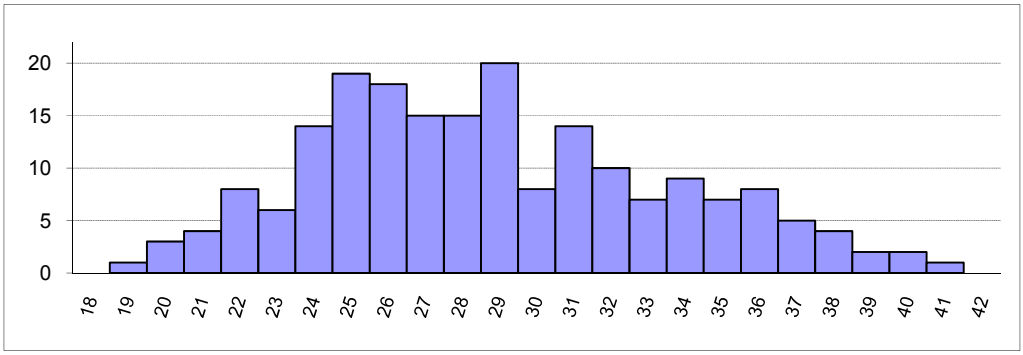
construíram-se 400 estimativas:

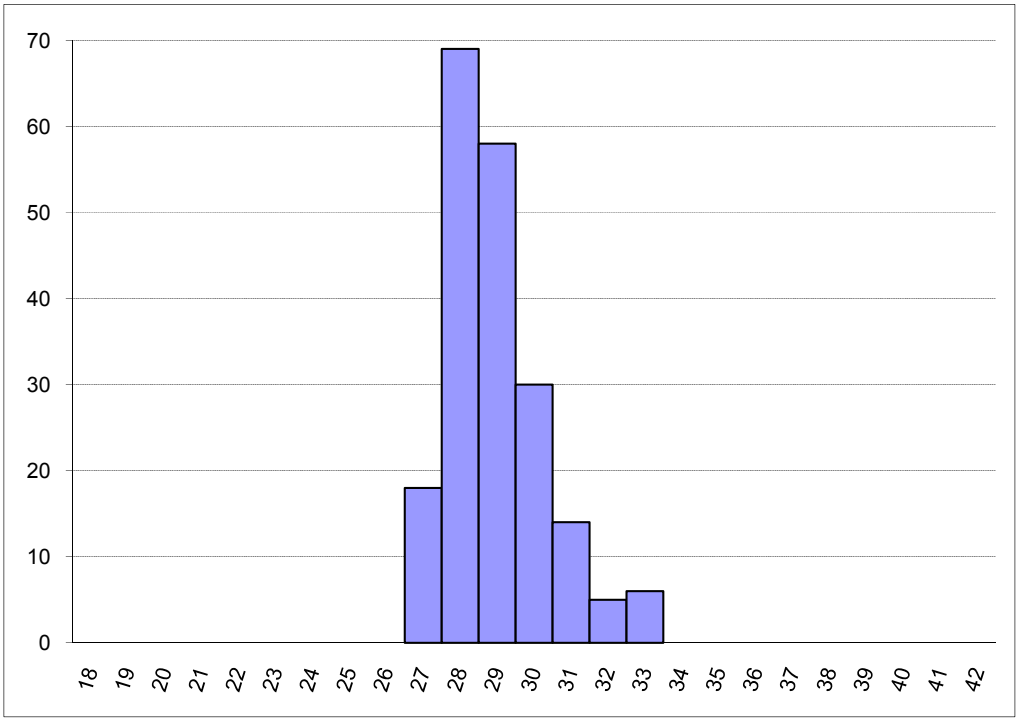
200 por índice

$$\hat{t}_{y_r} = \frac{\sum_{i=1}^{49} y_i}{\sum_{i=1}^{49} x_i} \times 22919$$

200 utilizando a média simples

$$\hat{t}_y = 196 * \bar{y} = 196 \times \frac{\sum_{i=1}^{49} y_i}{49}$$





4.3. Estimação em domínios ou subpopulações

Como obter, a partir de uma amostra aleatória simples de uma população, estimativas separadas para certos domínios ou subpopulações.

→ $(Y_1, Y_2, \ldots Y_i, \ldots, Y_n)$ amostra aleatória simples

destes n seleccionados existem n_d que pertencem ao domínio D e que vamos designar esse conjunto por S_d .

Pretende-se obter estimativas para a média (μ_d) ou para o total (t_{y_d}) da característica Y, para o conjunto dos N_d elementos do universo que pertencem ao domínio D, seja U_d .

Defina-se

$$U_i = \begin{cases} y_i & se \quad i \in U_d \\ 0 & se \quad i \notin U_d \end{cases} \qquad então \quad \sum_{i=1}^{n} U_i = \sum_{i \in S_d} Y_i \quad e \quad \sum_{s=1}^{N} U_s = \sum_{s \in U_d} y_s$$

$$X_i = \begin{cases} 1 & se \quad i \in U_d \\ 0 & se \quad i \notin U_d \end{cases} \qquad então \quad \sum_{i=1}^{n} X_i = n_d \quad e \quad \sum_{s=1}^{N} X_s = N_d$$

4.3.1. Estimador para a média do domínio e seu erro padrão

$$\mu_d = \frac{\sum_{i \in U_d} Y_i}{N_d} \quad \rightarrow \quad \textbf{estimador} \quad \rightarrow \quad \boxed{\hat{\mu}_d = \frac{\sum_{i \in S_d} Y_i}{n_d} = \bar{Y}_d} \quad = \frac{\sum_{i=1}^{n} U_i}{\sum_{i=1}^{n} X_i} = \hat{B}$$

Este estimador, que como se viu é a média dos valores da amostra que pertencem à subpopulação em estudo, pode ser encarado como o rácio entre duas variáveis e trata-se portanto da estimação de um quociente.

Demonstra-se, no entanto, que o enviesamento é praticamente nulo e assim, poder-se-á considerar que $\bar{Y}_d$ é aproximadamente um ***estimador centrado*** de μ_d .

O seu erro padrão

$$\boxed{EP(\hat{\mu}_d) \approx \sqrt{(1 - \frac{n}{N}) \frac{S'^2_d}{n_d}}} \qquad \text{onde} \quad S'^2_d = \frac{\sum_{i \in S_d} (Y_i - \bar{Y}_d)^2}{n_d - 1}$$

$$E\hat{Q}M(\hat{\mu}_d) \approx (1 - \frac{n}{N}) \frac{1}{n\bar{X}^2} \frac{\sum_{i=1}^{n} (U_i - \hat{B}X_i)^2}{n-1} = \qquad \text{como } \bar{X} = \frac{n_d}{n} \text{ e } \hat{B} = \bar{Y}_d$$

$$= (1 - \frac{n}{N}) \frac{1}{n(n_d/n)^2} \frac{\sum_{i \in S_d} (Y_i - \bar{Y}_d)^2}{n-1} = (1 - \frac{n}{N}) \frac{n}{n_d^2} \frac{(n_d - 1) S'^2_d}{n-1} \approx (1 - \frac{n}{N}) \frac{S'^2_d}{n_d}, \quad \text{pois } \frac{n_d - 1}{n-1} \approx \frac{n_d}{n}$$

4.3.2. Estimador para o total e seu erro padrão

I. **Conhece-se N_d,** nº de elementos do universo que pertencem à subpopulação.

Estimador $\hat{t}_{y_d} = N_d \overline{Y}_d$

Erro padrão $EP(\hat{t}_{y_d}) = N_d \; EP(\hat{\mu}_d)$

II. **Conhece-se o total** da característica para a população - t_y

Sendo $\hat{B}_d = \dfrac{\sum_{i=1}^{n} U_i}{\sum_{i=1}^{n} Y_i}$ (peso na amostra) e $U_i = \begin{cases} y_i & se \quad i \in U_d \\ 0 & se \quad i \notin U_d \end{cases}$

Estimador $\hat{t}_{y_d} = \hat{B}_d \, t_y$ estimador pelo quociente

Erro padrão $EP(\hat{t}_{y_d}) = t_y \sqrt{E\hat{Q}M(\hat{B}_d)}$

III. **Não se conhece** nem N_d nem t_y

Procede-se à extrapolação pura e simples

Estimador $\hat{t}_{y_d} = \dfrac{N}{n} \sum_{i \in S_d} Y_i = \dfrac{N}{n} \sum_{i=1}^{n} U_i = N\overline{U}$

Erro padrão $EP(\hat{t}_{y_d}) = N \sqrt{(1 - \dfrac{n}{N}) \dfrac{S'^2_U}{n}}$

onde $S'^2_U = \dfrac{1}{n-1}(\sum_{i=1}^{n} U_i^2 - n\overline{U}^2) = \dfrac{1}{n-1}(\sum_{i \in S_d} Y_i^2 - n\overline{U}^2)$, $\overline{U} = \dfrac{\sum_{i \in S_d} Y_i}{n}$

4.4. Estimação por regressão

Objectivo: aumentar a precisão dos estimadores utilizando uma variável auxiliar *X* que deverá estar correlacionada com a característica em estudo.

Hipóteses: conhecido o valor de μ_x ou t_x para a população;

bons resultados quando a regressão de Y sobre X é uma recta, ou seja, se $Y = B_o + B_1 X + \varepsilon$.

Para os ***n*** elementos que compõem a amostra aleatória simples observam-se os valores (X_i, Y_i).

4.4.1. Estimador de regressão e suas propriedades

O estimador de regressão linear para a ***média de Y*** ($\hat{\mu}_{y_{reg}}$).

$\hat{\mu}_{y_{reg}} = \hat{B}_o + \hat{B}_1 \mu_x$ onde $\hat{B}_o$ e $\hat{B}_1$ são os estimadores dos mínimos quadrados dos coeficientes do modelo de RLS, $y = B_o + B_1 x + \varepsilon$.

Assim, com $\hat{B}_o = \overline{Y} - \hat{B}_1 \overline{X}$ e $\hat{B}_1 = \dfrac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \dfrac{S'_{xy}}{S'^2_x} = \dfrac{r_{xy} S'_y}{S'_x}$,

$$\boxed{\hat{\mu}_{y_{reg}} = \overline{Y} + \hat{B}_1 (\mu_x - \overline{X})}$$ ***estimador por regressão***

Propriedades

Enviesamento: $E(\hat{\mu}_{y_{reg}}) = \mu_y - Cov(\hat{B}_1, \overline{X}) \Rightarrow Env(\hat{\mu}_{y_{reg}}) = -Cov(\hat{B}_1, \overline{X})$

Quanto maior for a associação linear entre *X* e Y, menor é o enviesamento, sendo mesmo *igual a zero* quando existe *relação linear perfeita*.

De qualquer modo

$\lim_{n\to\infty} Env(\hat{\mu}_{y_{reg}}) = \lim_{n\to\infty} -Cov(\hat{B}_1, \overline{X}) = 0$, e então $\hat{\mu}_{y_{reg}}$ é assintoticamente não enviesado.

Erro Quadrático Médio e sua estimação

$$EQM(\hat{\mu}_{y_{reg}}) \approx (1 - \frac{n}{N}) \frac{\sigma'^2_y}{n} (1 - \rho^2_{xy}) \qquad E\hat{Q}M(\hat{\mu}_{y_{reg}}) \approx (1 - \frac{n}{N}) \frac{s'^2_y}{n} (1 - r^2_{xy})$$

Esta estimativa para o EQM pode também ser calculada recorrendo à variância amostral dos resíduos. Assim pode também utilizar-se a estimativa

$E\hat{Q}M(\hat{\mu}_{y_{reg}}) \approx (1 - \frac{n}{N}) \frac{s'^2_e}{n}$ onde $s'^2_e = \dfrac{\sum_{i=1}^{n} e_i^2}{n-1}$, $e_i = y_i - \hat{y}_i$ resíduos dos M.Q.O.

Observações:

- Quanto ao enviesamento:
  - Estimadores geralmente enviesados mas assintoticamente não enviesados,
  - Enviesamento tanto menor quanto maior o grau de associação linear entre as variáveis. Sendo nulo se e só $\rho_{xy} = \pm 1$
- A precisão do estimador aumenta (ou seja o EQM diminui) quando:
  - aumenta a dimensão da amostra ($n \to \infty$),
  - aumenta a taxa de amostragem ($\frac{n}{N} \to 1$),
  - é grande o grau de associação linear entre X e Y ($\rho_{xy} \to \pm 1$)

4.4.2. Comparação da precisão dos estimadores

$$EQM(\overline{Y}) = V(\overline{Y}) = \frac{N-n}{Nn}\sigma_y'^2$$

$$EQM(\hat{\mu}_{y_r}) \approx \frac{N-n}{Nn}\left(\sigma_y'^2 + B^2\sigma_x'^2 - 2B\,\rho_{xy}\sigma_x'\sigma_y'\right)$$

$$EQM(\hat{\mu}_{y_{reg}}) \approx (1-\frac{n}{N})\frac{\sigma_y'^2}{n}(1-\rho_{xy}^2)$$

$$\Rightarrow \quad EQM(\hat{\mu}_{y_{reg}}) \le EQM(\overline{Y})$$

O estimador por regressão tem maior precisão que a média simples da amostra. A diferença será tanto maior quanto maior for a correlação entre as variáveis ($\rho_{xy}^2 \to 1$).

$$\Rightarrow \quad EQM(\hat{\mu}_{y_{reg}}) \le EQM(\hat{\mu}_{y_r})$$

Calculando

$$EQM(\hat{\mu}_{y_r}) - EQM(\hat{\mu}_{y_{reg}}) = \frac{N-n}{Nn}\left(\sigma_y'^2 + B^2\sigma_x'^2 - 2B\,\rho_{xy}\sigma_x'\sigma_y' - \sigma_y'^2 + \rho_{xy}^2\sigma_y'^2\right) =$$

$$= \frac{N-n}{Nn}\left(B\sigma_x' - \rho_{xy}\sigma_y'\right)^2 \qquad \ge 0$$

Sendo $EQM(\hat{\mu}_{y_r}) = EQM(\hat{\mu}_{y_{reg}}) \Leftrightarrow B\sigma_x' - \rho_{xy}\sigma_y' = 0 \Leftrightarrow B = \frac{\sigma_{xy}'}{\sigma_x^{2'}} = B_1$

isto é, a relação entre *Y* e *X* é representada por uma recta que passa pela origem.

4.4.3. Estimador pela diferença e suas propriedades

O estimador pela diferença é um caso particular do estimador pela regressão que pressupõe que a inclinação da recta de regressão é igual a 1.

$$\hat{\mu}_{y_{diff}} = \overline{Y} + (\mu_x - \overline{X})$$ ***estimador pela diferença***

$\Rightarrow \quad E(\hat{\mu}_{y_{diff}}) = \mu_y$ é não enviesado

PICR	**PISR**
$V(\hat{\mu}_{y_{diff}}) = \frac{\sigma_y^2 + \sigma_x^2 - 2\sigma_{xy}}{n}$	$(1-\frac{n}{N})\frac{\sigma_y'^2 + \sigma_x'^2 - 2\sigma_{xy}'}{n}$
$\hat{V}(\hat{\mu}_{y_{diff}}) = \frac{S_y'^2 + S_x'^2 - 2S_{xy}'}{n}$	$(1-\frac{n}{N})\frac{S_y'^2 + S_x'^2 - 2S_{xy}'}{n}$

Pode ainda generalizar-se este estimador para:

$$\hat{\mu}_{y_{diff}} = \overline{Y} + b_1(\mu_x - \overline{X})$$ com b_1, qualquer outro valor pré-determinado

4.5. Estimação por índice e por regressão em amostras estratificadas

4.5.1. Estimação por índice

Pretende-se estimar a média de *Y*, com base numa amostra estratificada, mas utilizando a informação conhecida sobre uma variável auxiliar *X*.

A. Estimadores separados ou independentes:

Para cada estrato h calcula-se a estimador por índice para a média ($\hat{\mu}_{yr_h}$)

$\hat{\mu}_{yr_h} = \frac{\overline{Y}_h}{\overline{X}_h}\mu_{x_h} = \hat{B}_h\,\mu_{x_h}$ estimador por índice para a média do estrato

O ***estimador por índice separado*** para a média da população obtém-se

$\hat{\mu}_{yr} = \sum_{h=1}^{H}\frac{N_h}{N}\hat{\mu}_{yr_h} = \sum_{h=1}^{H}W_h\hat{\mu}_{yr_h}$ média ponderada dos estimadores anteriores

Enviesamento:

$$E(\hat{\mu}_{yr}) = \sum_{h=1}^{H}\frac{N_h}{N}E(\hat{\mu}_{yr_h}) = \sum_{h=1}^{H}\frac{N_h}{N}\left[\mu_h + Env(\hat{\mu}_{y_h})\right] = \mu + \sum_{h=1}^{H}\frac{N_h}{N}Env(\hat{\mu}_{yr_h})$$

$Env(\hat{\mu}_{yr}) = \sum_{h=1}^{H}\frac{N_h}{N}Env(\hat{\mu}_{yr_h})$ média ponderada dos enviesamentos em cada estrato.

Estimador para o EQM:

$$E\hat{Q}M(\hat{\mu}_{y_r}) = \sum_{h=1}^{H}W_h^2 E\hat{Q}M(\hat{\mu}_{yr_h}) \approx \sum_{h=1}^{H}W_h^2(1-\frac{n_h}{N_h})\frac{1}{n_h}\left(S'^2_{Y_h} + \hat{B}_h{}^2 S'^2_{X_h} - 2\hat{B}_h\ S'_{XY_h}\right)$$

Observações:

- Em cada estrato têm de verificar-se as condições que tornam válida a utilização de estimadores por índice, nomeadamente, devem tratar-se de grandes amostras em cada estrato.
- Por outro lado também têm de ser conhecidos os valores das médias (ou dos totais) da variável auxiliar para cada estrato.

B. Estimador por índice combinado:

Com base numa amostra estratificada calculam-se os estimadores para a média da variável *X* e da variável *Y* : $\hat{\mu}_{YE} = \sum_{h=1}^{H}W_h\overline{Y}_h$ e $\hat{\mu}_{XE} = \sum_{h=1}^{H}W_h\overline{X}_h$

O **estimador por índice combinado** para a média de Y é dado por:

$$\hat{\mu}_{yr} = \frac{\hat{\mu}_{YE}}{\hat{\mu}_{XE}}\mu_X = \hat{B}_E\ \mu_X$$

estimador por índice construído a partir dos estimadores dos valores médios de *X* e *Y* produzidos com base numa amostra estratificada.

Estimador para o EQM:

$$E\hat{Q}M(\hat{\mu}_{y_r}) = \mu_X^2 E\hat{Q}M(\hat{B}_E) \approx \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \frac{1}{n_h} \left(S_{Y_h}^{'2} + \hat{B}_E^2 S_{X_h}^{'2} - 2\hat{B}_E \; S_{XY_h}^{'} \right)$$

Notas Finais:

Utilizar o 1º método, **ESTIMATIVAS SEPARADAS**:

- → se os B_h são variáveis de estrato para estrato;
- → desde que se verifiquem as condições para poder utilizar as estimativas por quociente para as médias dos estratos (nomeadamente os n_h serem suficientemente grandes para que o enviesamento tenda para 0 e se possam usar as aproximações do EQM);
- → sejam conhecidos os valores médios da característica X em cada estrato.

Utilizar o 2º método, **ESTIMATIVAS COMBINADAS**:

- → se os B_h são estáveis de estrato para estrato;
- → se as dimensões das amostras nos estratos são pequenas.

4.5.2. Estimador de regressão

A. Estimadores separados ou independentes:

Para cada estrato h calcula-se o estimador por regressão para a média de Y

$$\hat{\mu}_{yreg_h} = \overline{Y}_h + \hat{B}_{1_h}(\mu_{x_h} - \overline{X}_h) \quad \text{onde} \quad \hat{B}_{1_h} = \frac{\sum_{i=1}^{n_h} X_i Y_i - n_h \overline{X}_h \overline{Y}_h}{\sum_{i=1}^{n_h} X_{ih}^2 - n_h \overline{X}_h^2} = \frac{S_{XY_h}^{'}}{S_{X_h}^{'2}}$$

O estimador para a média da população

$$\hat{\mu}_{yreg} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\mu}_{yreg_h} = \sum_{h=1}^{H} W_h \hat{\mu}_{yreg_h}$$

Estimador para o EQM:

$$E\hat{Q}M(\hat{\mu}_{y_{reg}}) \approx \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \frac{s_{Y_h}^{'2}}{n_h} (1 - r_{XY_h}^2)$$

Observações:

⇒ Apropriada quando os os B_{1h} são variáveis de estrato para estrato;

⇒ Têm de se verificar as condições necessárias para poder utilizar as estimativas por regressão para as médias dos estratos (nomeadamente os n_h serem suficientemente grandes);

⇒ Conhecidos os valores médios da característica X em cada estrato.

B. Estimador combinado:

Com base numa amostra estratificada calculam-se os estimadores para a média da variável X e da variável Y : $\hat{\mu}_{YE} = \sum_{h=1}^{H} W_h \overline{Y}_h$ e $\hat{\mu}_{XE} = \sum_{h=1}^{H} W_h \overline{X}_h$

O estimador para a média de Y é dado por:

$$\hat{\mu}_{yreg} = \hat{\mu}_{YE} + \hat{B}_C(\mu_X - \hat{\mu}_{XE}) \text{ , } \quad \text{com} \quad \hat{B}_C = \frac{\sum_{h=1}^{H} W_h(1-\frac{n_h}{N_h})S'_{XY_h}}{\sum_{h=1}^{H} W_h(1-\frac{n_h}{N_h})S'^2_{X_h}}$$

Estimador para o EQM:

$$E\hat{Q}M(\hat{\mu}_{y_{reg}}) \approx \sum_{h=1}^{H} W_h^2(1-\frac{n_h}{N_h})\frac{1}{n_h}\left(S'^2_{Y_h} + \hat{B}_C^2 S'^2_{X_h} - 2\hat{B}_C\, S'_{XY_h}\right)$$

5. AMOSTRAGEM POR CONGLOMERADOS

Leitura obrigatória: pontos 5.1, 5.2 e 5.5 do capítulo 5 e pontos 6.1 e 6.2 do capítulo 6 do livro "Sampling: Design and Analysis", Sharon L. Lohr

As *unidades estatísticas* que compõem a população estão agrupadas em *unidades primárias (conglomerados)* que formam a base de sondagem donde é seleccionada a amostra.

Muito utilizada, por ex.: controlo de qualidade (artigos agrupados em lotes): a amostra final é composta por todos os artigos que pertencem aos lotes seleccionados para a amostra (unidades primárias).

5.1. Conceitos e notação

Seja então uma população com ***K unidades estatísticas*** dispostas do seguinte modo:

⇒ ***N unidades primárias*** *(U_1, U_2,...,U_N)* e em cada uma delas M_i unidades estatísticas, donde $M_1+M_2+...\ M_i+...+M_N=K$.

ou seja cada $U_i = \{u_{i1}, u_{i2},..., u_{iMi}\}$ com $i = 1, 2, ..., N$

Na população

N *unidades primárias (UP) - conglomerados*

	1	*2*	...	*i*	...	*N*	Total
Dimensão	M_1	M_2	...	M_i	...	M_N	K
Média	μ_1	μ_2	...	μ_i	...	μ_N	μ
Total	t_1	t_2	...	t_i	...	t_N	t_y
Variância	σ_1^2	σ_2^2	...	σ_i^2	...	σ_N^2	σ^2

μ_i e σ_i^2 (ou $\sigma_i'^2$) média e variância da característica na unidade primária i;

$\mu = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{y_{ij}}{K} = \sum_{i=1}^{N}\frac{M_i}{K}\mu_i$ média da característica na população;

$t_y = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N} M_i\mu_i = \sum_{i=1}^{N} t_i$ total da característica na população;

$$\sigma^2 = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{(y_{ij}-\mu)^2}{K} = \sigma_d^2 + \sigma_e^2,\ \sigma_d^2 = \sum_{i=1}^{N}\frac{M_i}{K}\sigma_i^2 \text{ e } \sigma_e^2 = \sum_{i=1}^{N}\frac{M_i}{K}(\mu_i-\mu)^2$$

Variância da população = soma da variância dentro das UP, σ_d^2 (média ponderada das variâncias de cada UP), e da variância entre as unidades primárias, σ_e^2.

Na amostra

Selecciona-se uma amostra aleatória de ***n*** conglomerados (*UP*)

taxa de amostragem

$U_{\alpha_1}, \ldots, U_{\alpha_i}, \ldots, U_{\alpha_n}$ $\quad f_1 = n/N$

e são observadas todas as m_i unidades finais que pertencem ao *i-ésimo* conglomerado seleccionado

***n** Conglomerados (UP)*

	U_{α_1}	U_{α_2}	...	U_{α_i}	...	U_{α_n}	Total
Dimensão	m_1	m_2	...	m_i	...	m_n	m
Média	$\overline{Y}_1$	$\overline{Y}_2$	...	$\overline{Y}_i$	...	$\overline{Y}_n$	$\overline{Y}$
Total	T_1	T_2	...	T_i	...	T_n	T
Variância	S_1^2	S_2^2	...	S_i^2	...	S_n^2	S^2
	$S_1'^2$	$S_2'^2$		$S_i'^2$		$S_n'^2$	S'^2

onde,

$T_i = \sum_{j=1}^{m_i} Y_{ij}$ total da característica na *i-ésima UP* da amostra

$\overline{Y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} Y_{ij} = \frac{T_i}{m_i}$ média da característica na *i-ésima UP* da amostra

$S_i^2 = \frac{1}{m_i} \sum_{j=1}^{m_i} (Y_{ij} - \overline{Y}_i)^2$ variância da característica na *i-ésima UP* da amostra

$\overline{Y} = \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{m_i} Y_{ij} = \frac{T}{m}$ média da amostra global, onde $m = \sum_{i=1}^{n} m_i$ e $T = \sum_{i=1}^{n} T_i$

$$S^2 = \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (Y_{ij} - \overline{Y})^2 = \frac{1}{m} \left[\sum_{i=1}^{n} \sum_{j=1}^{m_i} (Y_{ij} - \overline{Y}_i)^2 + \sum_{i=1}^{n} \sum_{j=1}^{m_i} (\overline{Y}_i - \overline{Y})^2 \right] =$$

$$= \frac{1}{m} \left[\sum_{i=1}^{m_i} m_i S_i^2 + \sum_{i=1}^{m_i} m_i (\overline{Y}_i - \overline{Y})^2 \right]$$ variância da amostra global

$\underbrace{\text{dentro das U.P.}}$ + $\underbrace{\text{entre U.P.}}$ = variação total

5.2. Conglomerados de igual dimensão

Considere-se o caso mais simples em que os ***conglomerados têm todos a mesma dimensão (M)***, ou seja,

$M_i = m_i = \boldsymbol{M} = K/N \qquad i = 1, 2, \ldots, N$

em que se selecciona uma **amostra aleatória simples de dimensão *n***.

5.2.1. Estimadores a utilizar e sua variância

Estima-se o total ou a média da população muito facilmente: trata-se a média ($\overline{Y}_i$) ou o total (T_i) de cada conglomerado como observações que são, de uma amostra aleatória simples de dimensão n.

⇒ Para o **total da população, t_y** Amostra aleatória simples ($T_1, T_2, ..., T_i, ..., T_n$)

Estimador

$$\boxed{\hat{T}_{Co} = \frac{N}{n}\sum_{i=1}^{n} T_i = N\overline{T}} \text{ com } \overline{T} = \frac{\sum_{i=1}^{n} T_i}{n} = \frac{T}{n}$$ Estimador não enviesado

Variância do estimador

	PICR	PISR
$V(\hat{T}_{Co}) = N^2 V(\overline{T}) =$	$N^2 \frac{\sigma_t^2}{n}$	$N^2(1-\frac{n}{N})\frac{\sigma_t'^2}{n}$

onde: $\sigma_t^2 = \frac{1}{N}\sum_{i=1}^{N}(t_i - \mu_t)^2$ e $\sigma_t'^2 = \frac{1}{N-1}\sum_{i=1}^{N}(t_i - \mu_t)^2$ e $\mu_t = \frac{t_y}{N}$

sua estimação

$\hat{V}(\hat{T}_{Co}) = N^2 \hat{V}(\overline{T}) =$	$N^2 \frac{S_t'^2}{n}$	$N^2(1-\frac{n}{N})\frac{S_t'^2}{n}$ onde $S_t'^2 = \frac{\sum_{i=1}^{n}(T_i - \overline{T})^2}{n-1}$

⇒ Para a **média da população, μ** Amostra aleatória simples ($\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_i, \ldots, \overline{Y}_n$)

Estimador

$$\hat{\mu}_{Co} = \frac{T}{Mn} = \frac{\sum_{i=1}^{n} T_i}{Mn} = \frac{1}{n}\frac{\sum_{i=1}^{n}\sum_{j=1}^{M} Y_{ij}}{M} = \frac{\sum_{i=1}^{n}\overline{Y}_i}{n} = \overline{\overline{Y}}$$ Estimador não enviesado

Variância do estimador

	PICR	**PISR**
$V(\hat{\mu}_{Co}) =$	$\frac{\sigma_t^2}{M^2 n} = \frac{\sigma_{\overline{y}}^2}{n}$	$(1-\frac{n}{N})\frac{\sigma_t'^2}{M^2 n} = (1-\frac{n}{N})\frac{\sigma_{\overline{y}}'^2}{n}$

onde $\sigma_{\overline{y}}^2 = \frac{\sum_{i=1}^{N}(\mu_i - \mu)^2}{N}$ e $\sigma_{\overline{y}}'^2 = \frac{\sum_{i=1}^{N}(\mu_i - \mu)^2}{N-1}$

sua estimação

	PICR	**PISR**
$\hat{V}(\hat{\mu}_{Co}) =$	$\frac{S_t'^2}{M^2 n} = \frac{S_{\overline{y}}'^2}{n}$	$(1-\frac{n}{N})\frac{S_t'^2}{M^2 n} = (1-\frac{n}{N})\frac{S_{\overline{y}}'^2}{n}$, $S_{\overline{y}}'^2 = \frac{\sum_{i=1}^{n}(\overline{Y}_i - \overline{\overline{Y}})^2}{n-1}$

5.2.2. Comparação com a amostragem aleatória simples

A amostragem por conglomerados quase sempre conduz a estimadores com menor precisão do que a obtida com uma amostra aleatória simples com o mesmo número de elementos.

Quadro da ANOVA – Amostragem por conglomerados

Fonte da variação	**Soma dos Quadrados [1]**	**g.l. [2]**	**Média da soma dos quadrados [1] / [2]**
Entre as UP	$SSB = \sum_{i=1}^{N}\sum_{j=1}^{M}(\mu_i-\mu)^2 = \sum_{i=1}^{N} M(\mu_i-\mu)^2$	*N-1*	$MSB=\sigma'^2_B$
Dentro das UP	$SSW = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu_i)^2$	*N(M-1)*	$MSW=\sigma'^2_W$
Total	$SST = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu)^2 = SSB+SSW$	*NM-1= = K-1*	σ'^2

$$V(\hat{\mu}_{Ca}) = (1-\frac{nM}{NM})\frac{\sigma'^2}{nM} = (1-\frac{n}{N})\frac{\sigma'^2}{nM} \qquad \textbf{[1]}$$

$$V(\hat{\mu}_{Co}) = (1-\frac{n}{N})\frac{\sigma'^2_{\bar{y}}}{n} = (1-\frac{n}{N})\frac{\sum_{i=1}^{N}(\mu_i-\mu)^2}{n(N-1)} = (1-\frac{n}{N})\frac{\sum_{i=1}^{N}M(\mu_i-\mu)^2}{nM(N-1)} =$$

$$= (1-\frac{n}{N})\frac{SSB}{nM(N-1)} = (1-\frac{n}{N})\frac{MSB}{nM} \qquad \textbf{[2]}$$

A $V(\hat{\mu}_{Co})$ será tanto maior quanto maior a variação entre conglomerados (*SSB*).
Comparando [1] e [2] poder-se-á afirmar que a amostragem aleatória simples terá maior precisão que a amostragem por conglomerados, isto é, $V(\hat{\mu}_{Co}) > V(\hat{\mu}_{Ca})$, quando $MSB > \sigma'^2$,ou seja, quando houver grande heterogeneidade entre os conglomerados.

5.2.3. Coeficiente de correlação intra-conglomerados **(ICC)**

O coeficiente de correlação intra-conglomerado é uma medida agregada da correlação entre os elementos do mesmo conglomerado. Obtém-se calculando o coeficiente de correlação de Pearson entre os *NM(M-1)* pares (y_{ij}, y_{ik}) (com $k \neq j$, $i = 1, 2, ..., N$), ou seja, ,

$$ICC = \frac{2\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=j+1}^{M}(y_{ij}-\mu)(y_{ik}-\mu)}{(M-1)(NM-1)\sigma'^2} = 1-\frac{M}{M-1}\times\frac{SSW}{SST}.$$

Como $0 \le \frac{SSW}{SST} \le 1$ então $-\frac{1}{M-1} \le ICC \le 1$.

É utilizado como uma medida do grau de homogeneidade dentro do conglomerado, pois:

- *ICC* é positivo se os elementos dentro dos conglomerados são semelhantes, neste caso *SSW* assume um valor baixo relativamente a *SST*, e o valor de *ICC* será elevado;

- ICC=1 quando SSW=0, ou seja quando há homogeneidade completa dentro do conglomerado;
- limite inferior de *ICC=-1/(M-1)* acontece quando $SSW \rightarrow SST$, ou seja quando $SSB \rightarrow 0$ o que significa **grande heterogeneidade dos elementos dentro** do conglomerado (conglomerados semelhantes entre si).

Por outro lado prova-se também que,

$$MSB = \frac{NM-1}{M(N-1)}\sigma'^2\left[1+(M-1)ICC\right]$$

e torna-se mais fácil responder à questão:

Quando é que a amostragem por conglomerados é mais precisa que a amostragem aleatória simples?

$$V(\hat{\mu}_{Co}) - V(\hat{\mu}_{Ca}) = (1-\frac{n}{N})\frac{MSB}{nM} - (1-\frac{n}{N})\frac{\sigma'^2}{nM} = (1-\frac{n}{N})\times\frac{\sigma'^2}{nM}\times\frac{M-1}{M(N-1)}\left[1+(NM-1)ICC\right]$$

$\Rightarrow$ A amostragem por conglomerados só terá maior precisão que a aleatória simples quando $-1/(M-1) \le ICC < -1/(NM-1)$, pois só nesta situação é que $V(\hat{\mu}_{Co}) < V(\hat{\mu}_{Ca})$, o que só ocorre quando é grande a heterogeneidade dentro dos conglomerados.

$\Rightarrow$ Desde que ***ICC* seja positivo**, o que ocorre na maior parte dos casos, a **amostragem por conglomerados tem pior precisão** que a aleatória simples; no entanto esta perda de precisão é muitas vezes compensada com a redução significativa de custos

$\Rightarrow$ Quanto **maior for ICC** tanto **maior é a perda de eficácia** da amostragem por conglomerados face à amostragem aleatória simples.

Só se pode definir o *ICC* quando os conglomerados têm dimensões iguais. Um indicador alternativo para medir a homogeneidade, e que se pode aplicar na generalidade das populações, é o coeficiente de determinação ajustado (R_a^2).

Variação Total = Variação dentro + Variação entre UP

$$SST = \sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij}-\mu)^2 = \sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij}-\mu_i)^2 + \sum_{i=1}^{N}M_i(\mu_i-\mu)^2 = SSW + SSB$$

$$R_a^2 = 1 - \frac{SSW/(K-N)}{SST/(K-1)} = 1 - \frac{MSW}{\sigma'^2}$$

Utiliza-se como medida de homogeneidade por causa da sua interpretação: parte da variabilidade total da característica explicada pelas diferenças entre as médias dos diversos conglomerados, ajustada pelos graus de liberdade.

Quando os conglomerados têm igual dimensão verifica-se que $R_a^2 \approx ICC$.

5.3. Conglomerados de diferentes dimensões

Quando se seleccionam aleatoriamente prédios de habitação para se inquirirem todos os seus agregados familiares, raramente se verificará um número de famílias igual em todos os prédios. De um modo geral os conglomerados não são de igual dimensão. A amostra é assim composta por *n* pares: (T_1, m_1), (T_2, m_2), …,(T_n, m_n)

5.3.1. Selecção com igual probabilidade

A. Estimador não enviesado

⇒ Para o **total da população, t_y**

Estima-se o total e a variância do estimador como anteriormente.

$$\hat{T}_{Co} = \frac{N}{n}\sum_{i=1}^{n} T_i = N\overline{T} \qquad \text{onde} \quad \overline{T} = \frac{\sum_{i=1}^{n} T_i}{n} = \frac{T}{n}$$

$$= \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} Y_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \omega_{ij} Y_{ij} \qquad \text{onde } \omega_{ij} = \frac{1}{\pi_{ij}} = \frac{N}{n}$$

Variância do estimador

	PICR	**PISR**
$V(\hat{T}_{Co}) = N^2 V(\overline{T}) =$	$N^2 \dfrac{\sigma_t^2}{n}$	$N^2(1-\dfrac{n}{N})\dfrac{\sigma_t'^2}{n}$

onde $\sigma_t^2 = \dfrac{1}{N}\sum_{i=1}^{N}(t_i - \mu_t)^2$ e $\sigma_t'^2 = \dfrac{1}{N-1}\sum_{i=1}^{N}(t_i - \mu_t)^2$

sua estimação

	PICR	**PISR**	
$\hat{V}(\hat{T}_{Co}) = N^2 \hat{V}(\overline{T}) =$	$N^2 \dfrac{S_t'^2}{n}$	$N^2(1-\dfrac{n}{N})\dfrac{S_t'^2}{n}$	onde $S_t'^2 = \dfrac{\sum_{i=1}^{n}(T_i - \overline{T})^2}{n-1}$

A diferença entre os dois casos é que: a variação entre os totais dos diferentes conglomerados (T_i) tenderá a aumentar quando as suas dimensões são diferentes. De um modo geral $S_t'^2$ é maior quando os conglomerados têm dimensões diferentes.

⇒ Para a **média da população, μ**

$$\hat{\mu}_{Co} = \frac{\hat{T}_{Co}}{K} = \frac{N}{nK}\sum_{i=1}^{n} T_i = \frac{N}{nK}\sum_{i=1}^{n}\sum_{j=1}^{m_i} Y_{ij} \qquad \text{Estimador não enviesado}$$

podendo ainda escrever-se,

$$\hat{\mu}_{Co} = \frac{1}{n}\sum_{i=1}^{n}\frac{m_i}{M_o}\overline{Y}_i \qquad \text{onde } M_o = \frac{K}{N} \text{ dimensão média de um conglomerado na população}$$

m_i dimensão do i-ésimo conglomerado seleccionado.

Para utilizar este estimador é ***necessário conhecer*** a dimensão total da população ***K***, número total de unidades finais $K = \sum_{i=1}^{N} M_i$, ou pelo menos a dimensão média por conglomerado na população, M_o.

Variância do estimador

	PICR	**PISR**
$V(\hat{\mu}_{Co}) =$	$\frac{N^2}{K^2}\frac{\sigma_t^2}{n}$	$\frac{N^2}{K^2}(1-\frac{n}{N})\frac{\sigma_t'^2}{n}$

sua estimação

	PICR	**PISR**
$\hat{V}(\hat{\mu}_{Co}) =$	$\frac{N^2}{K^2}\frac{S_t'^2}{n}$	$\frac{N^2}{K^2}(1-\frac{n}{N})\frac{S_t'^2}{n}$

podendo também utilizar-se

	PICR	**PISR**
$\hat{V}(\hat{\mu}_{Co}) =$	$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{m_i}{M_o}\overline{Y}_i - \hat{\mu}_{Co}\right)^2$	$\frac{1}{n}(1-\frac{n}{N})\sum_{i=1}^{n}\left(\frac{m_i}{M_o}\overline{Y}_i - \hat{\mu}_{Co}\right)^2$

B. Estimador de um rácio

Pode também pensar-se em μ_y como um quociente entre duas variáveis quantitativas, $\mu_y = \frac{\sum_{i=1}^{N} t_i}{\sum_{i=1}^{N} M_i} = \frac{\mu_t}{M_o} = B$ e utilizar o estimador estudado no capítulo 3,

$\hat{\mu}_{Co_r} = \frac{\sum_{i=1}^{n} T_i}{\sum_{i=1}^{n} m_i} = \frac{\overline{T}}{\overline{m}} = \hat{B}$ onde $\overline{T} = \frac{\sum_{i=1}^{n} T_i}{n}$ e $\overline{m} = \frac{\sum_{i=1}^{n} m_i}{n}$ (Estimador assintotica/não enviesado)

Erro quadrático médio:

	PICR	***PISR***
$EQM(\hat{\mu}_{Co_r}) =$	$\frac{\sigma_t^2 + B^2\sigma_m^2 - 2B\sigma_{tm}}{nM_o^2}$	$(1-\frac{n}{N})\frac{\sigma_t'^2 + B^2\sigma_m'^2 - 2B\sigma'_{tm}}{nM_o^2}$

que pode ser estimado:

$$E\hat{Q}M(\hat{\mu}_{Co_r}) \approx (1-\frac{n}{N})\frac{1}{nM_o^2}\sum_{i=1}^{n}\frac{(T_i - \hat{B}m_i)^2}{n-1} = (1-\frac{n}{N})\frac{1}{nM_o^2}\sum_{i=1}^{n}\frac{m_i^2(\overline{Y}_i - \hat{B})^2}{n-1}$$

$$\approx (1-\frac{n}{N})\frac{1}{nM_o^2}\frac{\sum_{i=1}^{n} T_i^2 + \hat{B}^2\sum_{i=1}^{n} m_i^2 - 2\hat{B}\sum_{i=1}^{n} T_i m_i}{n-1}$$

$$\approx (1-\frac{n}{N})\frac{1}{nM_o^2}\left(S_t'^2 + \hat{B}^2 S_m'^2 - 2\hat{B}S'_{tm}\right)$$

onde,

$$S_t'^2 = \frac{\sum_{i=1}^{n}(T_i - \overline{T})^2}{n-1},\ S_m'^2 = \frac{\sum_{i=1}^{n}(m_i - \overline{m})^2}{n-1} \text{ e } S'_{tm} = \frac{\sum_{i=1}^{n}(T_i - \overline{T})(m_i - \overline{m})}{n-1}$$

Este estimador é obrigatoriamente utilizado quando não se conhece nem K nem M_o. Neste caso para calcular o estimador para o EQM tem de substituir-se no denominador das expressões acima M_o por $\overline{m}$.

5.3.2. Selecção com probabilidades diferentes e com reposição

EXEMPLO:

Numa zona existem 4 armazéns com dimensões entre 100 e 1000 m^2. Pretende-se estimar as vendas totais dos armazéns da zona no último mês, seleccionando uma amostra de dimensão 1, com probabilidade proporcional à área do armazém.

POPULAÇÃO Armazéns	A	B	C	D	Total
Área (m^2)	100	200	300	1000	1600
Prob. de inclusão (ψ_i)	1/16	2/16	3/16	10/16	1
Vendas no último mês (t_i) *(característica a estimar)*	11	20	24	245	300

Para calcular a estimativa para o total de vendas utiliza-se a metodologia apresentada anteriormente, mesmo com diferentes probabilidades de inclusão. Cada observação tem associado um peso e, assim, a estimativa para o total tendo por base uma amostra de dimensão 1, obtém-se multiplicando o valor observado pelo peso dessa observação (peso que é igual ao inverso da probabilidade de selecção ($\omega_i = \frac{1}{\psi_i}$). Assim o estimador será

$\hat{T}_\psi = \frac{T_i}{\psi_i} = \omega_i T_i$.

Neste exemplo, é possível obter 4 amostras de dimensão 1:

AMOSTRAS possíveis Armazém seleccionado	A	B	C	D	Total
Vendas último mês (T_i)	11	20	24	245	300
Prob. de selecção	1/16	2/16	3/16	10/16	1
Estimativa para o total $\left(\hat{T}_\psi = \frac{T_i}{\psi_i}\right)$	176	160	128	392	

Assim calcula-se $E(\hat{T}_\psi)$ e $V(\hat{T}_\psi)$:

$E(\hat{T}_\psi) = 176 \times \frac{1}{16} + 160 \times \frac{2}{16} + 128 \times \frac{3}{16} + 392 \times \frac{10}{16} = 300 = t_y$ não enviesado

$V(\hat{T}_\psi) = (176-300)^2 \times \frac{1}{16} + (160-300)^2 \times \frac{2}{16} + (128-300)^2 \times \frac{3}{16} + (392-300)^2 \times \frac{10}{16}$ ***= 14 248***

Se a amostra fosse aleatória simples (ψ_i=1/4 para i=1,2,3,4)

AMOSTRAS possíveis Armazém seleccionado	A	B	C	D	Total
Vendas últ. mês (T_i)	11	20	24	245	300
Prob. de selecção	1/4	1/4	1/4	1/4	1
Estimativa para o total $\left(\hat{T}_y = \frac{T_i}{\pi_i}\right) = \frac{N}{n} T_i$	44	80	96	980	

$$E(\hat{T}_y) = \frac{44+80+96+980}{4} = 300 = t_y \quad \text{não enviesado}$$

$$V(\hat{T}_y) = \frac{(44-300)^2 + (80-300)^2 + (96-300)^2 + (980-300)^2}{4} = \mathbf{154\,488}$$

Como se pode ver ambas as metodologias conduzem a estimadores não enviesados, mas neste exemplo a selecção PICR é muito menos eficiente que a selecção com probabilidade variável.

A) <u>Caso Geral</u>

Seja então uma amostra de n conglomerados seleccionados com probabilidades diferentes mas com reposição.

Isto significa que em cada uma das tiragens pode ser seleccionada cada uma das UP da população com probabilidades que, embora diferentes ($\psi_1, \psi_2, \ldots \psi_N$), não variam de tiragem para tiragem uma vez que a selecção é com reposição.

A ideia subjacente à amostragem com diferentes probabilidades e com reposição é simples e resume-se ao seguinte:

⇒ Seleccionar n conglomerados, com diferentes probabilidades mas com reposição, isto é, em cada uma das n tiragens tem-se uma v.a. T_i com distribuição dada por:

	t_1	t_2	...	t_i	...	t_N
Probabilidade	ψ_1	ψ_2	...	ψ_i	...	ψ_N

⇒ Estimar o total da população com base na informação fornecida por cada elemento da amostra, obtendo assim ***n* estimadores para o total**, que em virtude da selecção ser com reposição são variáveis aleatórias *iid*, isto é,

$$\hat{T}_i = \frac{T_i}{\psi_i} \qquad i=1, 2, \ldots, n$$

⇒ Construir o estimador para o total pretendido com base na média simples dos estimadores anteriores, ou seja:

$$\hat{T}_\psi = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\psi_i} = \frac{1}{n}\sum_{i=1}^{N} Q_i \frac{t_i}{\psi_i}$$

Estimador centrado, $E(\hat{T}_\psi) = t_y$, onde Q_i = nº de vezes que U_i é seleccionado para a amostra

com variância

$$V(\hat{T}_{\psi}) = \frac{1}{n}\sum_{i=1}^{N}\psi_i\left(\frac{t_i}{\psi_i} - t_y\right)^2$$

$\Rightarrow$ A variância deste estimador é estimada através da expressão:

$$\hat{V}(\hat{T}_{\psi}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{T_i}{\psi_i} - \hat{T}_{\psi}\right)^2 \quad = \quad \frac{1}{n(n-1)}\sum_{i=1}^{N} Q_i\left(\frac{t_i}{\psi_i} - \hat{T}_{\psi}\right)^2$$

que se prova ser um estimador não enviesado de $V(\hat{T}_{\psi})$.

B) <u>Probabilidade proporcional à dimensão</u> (pps)

$\psi_i = \frac{M_i}{K} \quad i = 1,2,...,N$ probabilidade de selecção proporcional ao número de elementos do conglomerado no total da população.

	$TOTAL, t_y \quad (\hat{T}_{\psi})$	$MÉDIA, \mu \quad (\hat{\mu}_{\psi})$
Estimador para:	$\frac{K}{n}\sum_{i=1}^{n}\overline{Y}_i = K\overline{\overline{Y}}$	$\frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i = \overline{\overline{Y}}$
Variância do estimador	$\frac{K}{n}\sum_{i=1}^{N} M_i(\mu_i - \mu)^2 = K^2\frac{\sigma_e^2}{n}$	$\frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{K}(\mu_i - \mu)^2 = \frac{\sigma_e^2}{n}$
Erro Padrão	$\sqrt{\frac{K^2}{n(n-1)}\sum_{i=1}^{n}\left(\overline{Y}_i - \overline{\overline{Y}}\right)^2}$	$\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\overline{Y}_i - \overline{\overline{Y}}\right)^2}$

A selecção dos conglomerados com probabilidade proporcional à dimensão leva a que os conglomerados com maiores dimensões tenham maior probabilidade de pertencer à amostra.

Fundamental desenhar o mecanismo de selecção de forma a garantir a selecção dos conglomerados com a probabilidade desejada.

Dois métodos para seleccionar conglomerados com probabilidades diferentes (páginas 185 a 187):

- Cumulativo
- Método de Lahiri

5.4. Amostragem Sistemática

População com K elementos: $y_1, y_2, \ldots, y_i, \ldots, y_N, y_{N+1}, \ldots, y_K$ com média μ e variância σ^2.

1	2		i		N
y_1	y_2	$\cdots$	y_i	$\cdots$	y_N
			y_{i1}		
y_{N+1}	y_{N+2}	$\cdots$	y_{N+i}	$\cdots$	y_{2N}
			y_{i2}		
$\cdots$					
$Y_{(M-1)N+1}$	$y_{(M-1)N+2}$	$\cdots$	$y_{(M-1)N+i}$	$\cdots$	y_K
			y_{iM}		

$$N = \left[\frac{K}{M}\right]$$

Intervalo amostral

$K=MN$

$$\mu_i = \frac{\sum_{j=1}^{M} y_{ij}}{M} = \bar{y}_i$$

Selecção de uma amostra sistemática de dimensão M, genericamente, selecção de ***um*** conglomerado dos ***N*** existentes, (conglomerado ***i*** seleccionado com probabilidade *1/N)*:

y_i, y_{N+i}, y_{2N+i}, …, $y_{N(M-1)+I}$ *i-ésimo conglomerado*

$(y_{i1}$, y_{i2}, y_{ij}, …, $y_{iM})$

5.4.1. Estimadores a utilizar

Para a média ***Para o total***

$$\hat{\mu}_{sis} = \bar{Y}_i = \frac{\sum_{j=1}^{M} Y_{ij}}{M} \qquad \hat{T}_y = K\bar{Y}_i$$

$\Rightarrow$ Quando $K=MN$ tem-se que $E(\bar{Y}_i) = \mu$, estimador é não enviesado.

Variância do estimador:

$$\sigma^2_{\bar{Y}_i} = V(\hat{\mu}_{sis}) = \frac{\sum_{i=1}^{N} (\bar{y}_i - \mu)^2}{N} = \frac{1}{N}\frac{SSB}{M} = (1-\frac{1}{N})\frac{MSB}{M} \text{ pois } SSB = MSB \times (N-1)$$

como $MSB = \frac{N}{(N-1)}\sigma^2\left[1+(M-1)ICC\right]$ também se pode escrever: $\sigma^2_{\bar{Y}_i} = \frac{\sigma^2}{M}\left[1+(M-1)ICC\right]$,

onde ICC - Coeficiente de correlação intra-conglomerados conforme definido na amostragem por conglomerados de igual dimensão, com $-\frac{1}{M-1} \le ICC \le 1$.

Então,

$\Rightarrow$ Quando ***ICC= 0*** temos $\sigma^2_{\bar{Y}_i} = \frac{\sigma^2}{M}$, isto é, a variância do estimador da amostragem sistemática é idêntica a amostra ***PICR***.

$\Rightarrow$ Quando ***ICC= -1/(K-1)*** temos $\sigma^2_{\overline{Y}_i} = \frac{K-M}{K-1}\frac{\sigma^2}{M}$, isto é, a variância do estimador da amostragem sistemática é idêntica a amostra **PISR**.

Ocorre quando a ordem dos elementos da população é perfeitamente aleatória.

$\Rightarrow$ Quando ***ICC= 1*** a variância do estimador da amostragem sistemática é elevada.

Ocorre usualmente quando os elementos da população estão ordenados numa lista onde existe periodicidade no seu posicionamento. Grande homogeneidade dentro do conglomerado.

$\Rightarrow$ Quando $-\frac{1}{M-1} \le ICC \le -\frac{1}{K-1}$ então a amostra sistemática proporciona estimadores com maior precisão que uma amostra **PISR**.

Ocorre quando os elementos da população estão listados por ordem crescente ou decrescente da variável em estudo, ou de outra muito correlacionada com a variável em estudo.

EXEMPLO:

População consoante o número de ordem da visita

Nº Visita	1	2	3	4	5	6	7	8	9	10	11	12
Tempo	15	34	35	36	11	17	49	40	25	46	33	14

K=12 $\quad \mu_y = 29{,}583 \qquad \sigma^2_y = 153{,}08$

Se PISR $\sigma^2_{\overline{Y}} = \frac{K-M}{K-1}\frac{\sigma^2_y}{M} = \frac{12-3}{11}\times\frac{153{,}08}{3}$ **=41,75**

Amostra sistemática de 1 em cada 4 (N=4, M=3).

Amostras possíveis:

i	Visita	Tempos (y_{i1}, y_{i2}, y_{i3})	$\overline{Y}_i$	
			$\overline{y}_i$	Prob
1	1,5,9	15,11,25	17	¼
2	2,6,10	34,17,46	32,(3)	¼
3	3,7,11	35,49,33	39	¼
4	4,8,12	36,40,14	30	¼

$E(\overline{Y}_i) = \frac{1}{4}(17+32{,}3+39+30) = 29{,}583 = \mu_y$, $V(\overline{Y}_i) = \frac{1}{4}\left[(17-29{,}583)^2 + \ldots + (30-29{,}583)^2\right]$ **=63,7**

$$ICC = \frac{2\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=j+1}^{M}(y_{ij}-\mu_y)(y_{ik}-\mu_y)}{K\times(M-1)\times\sigma_y^2} = \frac{2\sum_{i=1}^{4}\sum_{j=1}^{3}\sum_{k=j+1}^{3}(y_{ij}-29{,}583)(y_{ik}-29{,}583)}{12\times 2\times 153{,}08} \textbf{=0,124}$$

Ordenando as observações por ordem crescente

Nº ordem	1	2	3	4	5	6	7	8	9	10	11	12
Tempo	11	14	15	17	25	33	34	35	36	40	46	49

$\mu_y = 29{,}583 \quad \sigma_y^2 = 153{,}08$

Amostras possíveis:

i	Nº ordem	Tempos (y_{i1}, y_{i2}, y_{i3})	$\bar{Y}_i$	
			$\bar{y}_i$	Prob
1	1,5,9	11,25,36	24	¼
2	2,6,10	14,33,40	29	¼
3	3,7,11	15,34,46	31,67	¼
4	4,8,12	17,35,49	33,67	¼

$E(\bar{Y}_i) = 29{,}583 = \mu_y \qquad V(\bar{Y}_i) = \frac{1}{4}\left[(24-29{,}583)^2 + \ldots + (33{,}67-29{,}583)^2\right] = \mathbf{13{,}13} \quad ICC = \mathbf{-0{,}371}$

Ordenando as observações de forma a haver periodicidade

Nº ordem	1	2	3	4	5	6	7	8	9	10	11	12
Tempo	11	17	34	40	14	25	35	46	15	33	36	49

$\mu_y = 29{,}583 \quad \sigma_y^2 = 153{,}08$

Amostras possíveis:

i	Nº ordem	Tempos (y_{i1}, y_{i2}, y_{i3})	$\bar{Y}_i$	
			$\bar{y}_i$	Prob
1	1,5,9	11,14,15	13,(3)	¼
2	2,6,10	17,25,33	25	¼
3	3,7,11	34,35,36	35	¼
4	4,8,12	40,46,49	45	¼

$E(\bar{Y}_i) = 29{,}583 = \mu_y \qquad V(\bar{Y}_i) = \frac{1}{4}\left[(13{,}3-29{,}583)^2 + \ldots + (45-29{,}583)^2\right] = \mathbf{138{,}02} \quad ICC = \mathbf{0{,}852}$

5.4.2. Estimação da variância do estimador

$\Rightarrow$ Como obter estimativas para $\sigma^2_{\bar{Y}_i} = V(\bar{Y}_i) = \dfrac{\sum_{i=1}^{N}(\bar{y}_i - \mu)^2}{N}$?

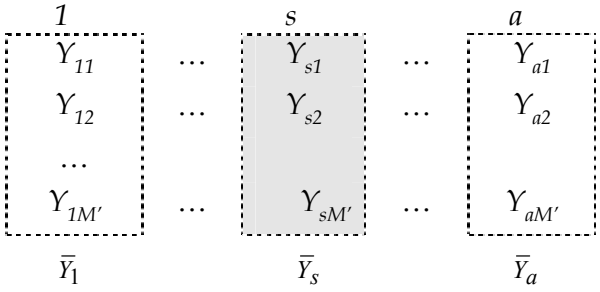
Muitas vezes, na prática, parte-se do princípio que a lista que serve de base à selecção da amostra sistemática está arrumada de forma aleatória assumindo equivalência à PISR.

Assim: $\hat{V}(\bar{Y}_i) \approx \dfrac{S'^2_Y}{M}\left(1 - \dfrac{M}{K}\right)$

No entanto se houver periodicidade $V(\bar{Y}_i)$ pode ser muito alta e esta estimativa não será válida.

A ***amostragem sistemática repetida*** conduz a estimativas para a variância dos estimadores que são não enviesadas qualquer que seja a ordem ou a periodicidade existentes.

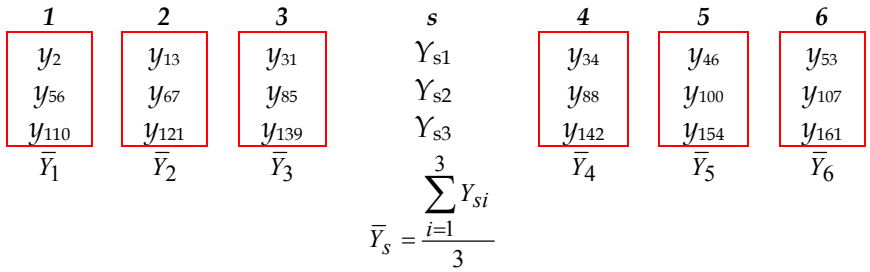
- Pretende-se uma amostra de M elementos dos K que formam a população.
- Então, em vez de seleccionar uma amostra sistemática de ***M*** elementos, seleccionam-se ***a*** amostras sistemáticas de ***M'*** elementos das ***A*** possíveis $\left(\frac{K}{M'} = A\right)$ e assim: $M = a \times M'$.



- Estimadores a utilizar :

$$\hat{\mu} = \frac{\sum_{s=1}^{a} \bar{Y}_s}{a} \qquad \hat{V}(\hat{\mu}) = \left(1 - \frac{a}{A}\right) \times \frac{1}{a} \times \frac{\sum_{s=1}^{a} (\bar{Y}_s - \hat{\mu})^2}{a-1}$$

EXEMPLO:
Pretende-se uma amostra sistemática de 18 elementos (M=18) dos 162 que compõem a população (K=162).
Então procede-se da seguinte forma: seleccionam-se 6 amostras sistemáticas de 3 elementos, $M = 18 = 6 \times 3$ ($a = 6$, $M' = 3$), das 54 possíveis ($A = \frac{162}{3} = 54$).
Como?
Por exemplo, gerando 6 numeros aleatórios entre 1 e 54. Suponhamos que os números foram o 2, 13, 46, 31, 34 e 53, e escolhem-se (a=) 6 amostras de amostras de (M'=)3 elementos,



$$\hat{\mu} = \frac{\sum_{s=1}^{6} \bar{Y}_s}{6} \quad \text{e} \quad \hat{V}(\hat{\mu}) = \left(1 - \frac{6}{54}\right) \times \frac{1}{6} \times \frac{\sum_{s=1}^{6} (\bar{Y}_s - \hat{\mu})^2}{5}$$

Quando *K/M* não é inteiro, $\overline{Y}_i$ não é estimador centrado para μ, embora o enviesamento tenda para zero quando *K* e *M* são razoavelmente grandes.

Com a metodologia seguinte obtêm-se sempre estimadores não enviesados:

1. Gerar um número aleatório entre *1* e *K* (número de elementos da população), designe-se por *j* ;
2. Calcular o quociente *j/N*, onde *N* é o intervalo amostral ($N = \left[\frac{K}{M}\right]$), e expressar esse quociente como um inteiro e um resto *r* ($\frac{23}{6} = 3 + \frac{5}{6}$, *r=5*);
3. Se o resto é zero tomar uma amostra sistemática de *1* em *N* elementos começando com o elemento *N*;

   Se o resto é ≠ de zero, *r*≠0, então tomar uma amostra sistemática de *1* em *N* elementos começando com o elemento *r*.

EXEMPLO: $K = 10$ $M = 3$ $N = \left[\frac{10}{3}\right] = 3$

Nº aleatório	***j/N***	**Resto (*r*)**	**Elementos na amostra**
1	1/3	1	1, 4, 7, 10
2	2/3	2	2, 5, 8
3	3/3	0	3, 6, 9
4	4/3	1	1, 4, 7, 10
5	5/3	2	2, 5, 8
6	6/3	0	3, 6, 9
7	7/3	1	1, 4, 7, 10
8	8/3	2	2, 5, 8
9	9/3	0	3, 6, 9
10	10/3	1	1, 4, 7, 10

3 amostras distintas:

Am 1. (1, 4, 7, 10)

$\overline{y}_1 = (y_1 + y_4 + y_7 + y_{10})/4$

Am 2. (2, 5, 8)

$\overline{y}_2 = (y_2 + y_5 + y_8)/3$

Am 3. (3, 6, 9)

$\overline{y}_3 = (y_3 + y_6 + y_9)/3$

A distribuição de $\overline{Y}_i$ é dada por:

$\overline{Y}_i$	$\overline{y}_1$	$\overline{y}_2$	$\overline{y}_3$
$f(\overline{y}_i)$	4/10	3/10	3/10

$$E(\overline{Y}_i) = \sum_{i=1}^{3} \overline{y}_i \times f(\overline{y}_i) = \overline{y}_1 \times \frac{4}{10} + \overline{y}_2 \times \frac{3}{10} + \overline{y}_3 \times \frac{3}{10} = \frac{\sum_{j=1}^{10} y_i}{10} = \mu$$

6. AMOSTRAGEM BI-ETÁPICA

Leitura obrigatória: pontos 5.3, 5.4 do capítulo 5 e ponto 6.3 do capítulo 6 do livro "Sampling: Design and Analysis", Sharon L. Lohr

A *amostragem bi-etápica* é um caso particular da amostragem por etapas múltiplas.

As *unidades estatísticas* que compõem a população estão agrupadas em *unidades primárias* que formam a base de sondagem donde é seleccionada a amostra.

1ª etapa: seleccionam-se algumas unidades primárias (por exemplo: turmas)

2ª etapa: seleccionam-se algumas unidades estatísticas (alunos) dentro das *UP* (turmas) seleccionadas anteriormente.

6.1. Conceitos e notação

Seja então uma população com ***K unidades estatísticas*** dispostas do seguinte modo:

$\Rightarrow$ ***N unidades primárias*** $(U_1, U_2,...,U_N)$ e em cada uma delas M_i unidades estatísticas, donde $M_1+M_2+...\ M_i+...+M_N=K$.

ou seja cada $U_i = \{u_{i1}, u_{i2},..., u_{iMi}\}$ com $i = 1, 2, ..., N$

Na população

N *unidades primárias (UP)*

	1	*2*	...	*i* $\{u_{i1}, u_{i2},..., u_{iMi}\}$	...	*N*	Total
Dimensão	M_1	M_2	...	M_i	...	M_N	K
Média	μ_1	μ_2	...	μ_i	...	μ_N	μ
Total	t_1	t_2	...	t_i	...	t_N	t_y
Variância	σ_1^2	σ_2^2	...	σ_i^2	...	σ_N^2	σ^2

μ_i e σ_i^2 (ou σ_i^2) média e variância da característica na unidade primária i;

$\mu = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{y_{ij}}{K} = \sum_{i=1}^{m}\frac{M_i}{M}\mu_i$ média da característica na população;

$t_y = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N} M_i\mu_i = \sum_{i=1}^{N} t_i$ total da característica na população;

$$\sigma^2 = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{(y_{ij}-\mu)^2}{K} = \sigma_d^2 + \sigma_e^2,\ \sigma_d^2 = \sum_{i=1}^{N}\frac{M_i}{K}\sigma_i^2 \text{ e } \sigma_e^2 = \sum_{i=1}^{N}\frac{M_i}{K}(\mu_i-\mu)^2$$

Variância da população = soma da variância dentro das UP, σ_d^2 (média ponderada das variâncias de cada UP), e da variância entre as unidades primárias, σ_e^2.

Na amostra

1ª etapa

Selecciona-se uma amostra aleatória de ***n*** unidades primárias das ***N*** existentes na população

taxa de amostragem

$U_{\alpha_1}, \ldots, U_{\alpha_i}, \ldots, U_{\alpha_n}$ $\quad f_1 = n/N$

2ª etapa

De cada *UP* seleccionada na 1ª etapa serão seleccionadas amostras aleatórias simples de dimensão m_i unidades finais das M_i existentes

Da UP U_{α_i} taxa de amostragem

$Y_{1\alpha_i}, \ldots, Y_{j\alpha_i}, \ldots, Y_{m_i\alpha_i}$ $\quad f_{2i} = m_i / M_i$

***n** unidades primárias*

	U_{α_1}	U_{α_2}	...	U_{α_i}	...	U_{α_n}	Total
Dimensão	m_1	m_2	...	m_i	...	m_n	m
Média	$\overline{Y}_1$	$\overline{Y}_2$	...	$\overline{Y}_i$	...	$\overline{Y}_n$	$\overline{Y}$
Total	T_1	T_2	...	T_i	...	T_n	T
Variância	$S_1'^2$	$S_2'^2$	...	$S_i'^2$	...	$S_n'^2$	S'^2

$T_i = \sum_{j=1}^{m_i} Y_{ij}$ total das observações da *i-ésima UP* da amostra

$\overline{Y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} Y_{ij} = \frac{T_i}{m_i}$ média das observações da *i-ésima UP* da amostra

$S_i^2 = \frac{1}{m_i} \sum_{j=1}^{m_i} (Y_{ij} - \overline{Y}_i)^2$ variância das observações da *i-ésima UP* da amostra

$\overline{Y} = \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{m_i} Y_{ij} = \frac{T}{m}$ média da amostra global, onde $m = \sum_{i=1}^{n} m_i$ e $T = \sum_{i=1}^{n} T_i$

$$S^2 = \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (Y_{ij} - \overline{Y})^2 = \frac{1}{m} \left[\sum_{i=1}^{n} \sum_{j=1}^{m_i} (Y_{ij} - \overline{Y}_i)^2 + \sum_{i=1}^{n} \sum_{j=1}^{m_i} (\overline{Y}_i - \overline{Y})^2 \right] =$$

$$= \frac{1}{m} \left[\sum_{i=1}^{n} m_i S_i^2 + \sum_{i=1}^{n} m_i (\overline{Y}_i - \overline{Y})^2 \right]$$ variância da amostra global

$\underbrace{\text{dentro das U.P.}}$ + $\underbrace{\text{entre U.P.}}$ = variação total

6.2. Estimadores a utilizar

Para cada unidade primária seleccionada, U_{α_i} , a estimação de μ_i (ou de t_i) , é feita utilizando $\hat{\mu}_i = \overline{Y}_i$ (ou $\hat{T}_i = M_i \overline{Y}_i$), que se sabe ser estimador não enviesado e consistente, com variância dada por $\frac{\sigma_i^2}{m_i}$ (ou $M_i^2 \frac{\sigma_i^2}{m_i}$) ou $\frac{M_i - m_i}{M_i} \frac{\sigma_i'^2}{m_i}$ (ou $\frac{M_i(M_i - m_i)\sigma_i'^2}{m_i}$), consoante a tiragem seja com ou sem reposição.

Para estimar a média da população vamos utilizar uma combinação linear das médias amostrais de cada uma das *n* UP, ou seja

$$\hat{\mu}_{Bi} = \sum_{i=1}^{n} a_i \overline{Y}_i \qquad \text{onde} \quad a_i : E(\hat{\mu}_{Bi}) = \mu$$

$$E(\hat{\mu}_{Bi}) = E(\sum_{i=1}^{n} a_i \overline{Y}_i) = \sum_{i=1}^{n} E(a_i \overline{Y}_i) = \mu$$

Mas como na 1ª etapa a α_i-ésima unidade primária da amostra, $\boldsymbol{U_{\alpha_i}}$, pode ser qualquer uma das $\boldsymbol{N}$ unidades primárias $\boldsymbol{U_i}$ pertencentes à população, cada uma com probabilidade ψ_i, tem-se que:

$$E(a_i \overline{Y}_i) = \psi_1 a_1 E(\overline{Y}_1) + \psi_2 a_2 E(\overline{Y}_2) + \ldots + \psi_N a_N E(\overline{Y}_N) = \sum_{i=1}^{N} \psi_i a_i \mu_i$$

$$E(\hat{\mu}_{Bi}) = \sum_{i=1}^{n} E(a_i \overline{Y}_i) = n \sum_{i=1}^{N} \psi_i a_i \mu_i \quad \text{e como } \mu = \sum_{i=1}^{N} \frac{M_i}{K} \mu_i \ ,$$

para garantir o não enviesamento: $a_i = \frac{1}{n} \frac{M_i}{\psi_i K}$ para $i = 1, 2, \ldots, N$

Estimadores não enviesados e consistentes:

Média da população	***Total da população***
$\hat{\mu}_{Bi} = \frac{1}{n} \sum_{i=1}^{n} \frac{M_i}{K \psi_i} \overline{Y}_i$	$\hat{T}_{Bi} = \frac{1}{n} \sum_{i=1}^{n} \frac{M_i}{\psi_i} \overline{Y}_i = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{T}_i}{\psi_i}$

dependendo somente da probabilidade de selecção das ***UP*** (1ª etapa).

Variância dos estimadores:

Se a ***selecção das UP for feita com reposição***, então existe independência entre as tiragens da 1ª etapa e prova-se que

$$V(\hat{\mu}_{Bi}) = \frac{1}{n^2}\sum_{i=1}^{n} V\left(\frac{M_i}{K\,\psi_i}\overline{Y}_i\right) = \frac{1}{n}\left\{\sum_{i=1}^{N}\frac{M_i^2}{K^2\,\psi_i}\left[V(\overline{Y}_i)+\mu_i^2\right] - \mu^2\right\}$$

$$V(\hat{T}_{Bi}) = \frac{K^2}{n^2}\sum_{i=1}^{n} V\left(\frac{M_i}{K\,\psi_i}\overline{Y}_i\right) = \frac{1}{n}\left\{\sum_{i=1}^{N}\frac{M_i^2}{\psi_i}\left[V(\overline{Y}_i)+\mu_i^2\right] - t_y^2\right\}$$

Estimador para a variância dos estimadores

Se a selecção na 1ª etapa for **com reposição** (ou mesmo sem reposição se a taxa de amostragem na 1ª etapa for baixa)

$$\hat{V}(\hat{\mu}_{Bi}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{Mi}{K\,\psi_i}\overline{Y}_i - \hat{\mu}_{Bi}\right)^2$$

$$\hat{V}(\hat{T}_{Bi}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{Mi}{\psi_i}\overline{Y}_i - \hat{T}_{Bi}\right)^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{\hat{T}_i}{\psi_i} - \hat{T}_{Bi}\right)^2$$

são estimadores não enviesados e consistentes para as variâncias dos estimadores)

6.2.1. Selecção das UP com reposição

6.2.1.1 Probabilidade de selecção das UP proporcional à dimensão (pps)

Se, na 1ª etapa, as unidades primárias são seleccionadas com probabilidade proporcional à sua dimensão: $\psi_i = \frac{M_i}{K}$ $\quad i = 1, 2, \ldots, N$

Estimadores não enviesados

$$\boxed{\hat{\mu}_{Bi} = \frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i = \overline{\overline{Y}}} \qquad \boxed{\hat{T}_{Bi} = \frac{K}{n}\sum_{i=1}^{n}\overline{Y}_i = K\overline{\overline{Y}}} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{K}{nm_i}Y_{ij}$$

Se a amostra da 1ª etapa é seleccionada com reposição e probabilidade proporcional à dimensão e na 2ª etapa a amostra é aleatória simples sem reposição, o peso amostral (inverso da probabilidade de selecção) de cada elemento da amostra pertencente à **UP$_i$** ***é*** $\omega_i = \frac{K}{nm_i}$**.**

Variância dos estimadores

$$V(\hat{\mu}_{Bi}) = \frac{1}{n}\left\{\sum_{i=1}^{N}\frac{M_i}{K}\left[V(\overline{Y}_i)+\mu_i^2\right]-\mu^2\right\} = \frac{1}{n}\left[\sum_{i=1}^{N}\frac{M_i}{K}\ V(\overline{Y}_i)+\sigma_e^2\right]$$

$$V(\hat{T}_{Bi}) = \frac{K^2}{n}\left[\sum_{i=1}^{N}\frac{M_i}{K}\ V(\overline{Y}_i)+\sigma_e^2\right] \quad \text{com} \quad \sigma_e^2 = \sum_{i=1}^{m}\frac{M_i}{M}(\mu_i-\mu)^2$$

Estimadores para estas variâncias

$$\hat{V}(\hat{\mu}_{Bi}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\overline{Y}_i-\overline{\overline{Y}})^2 = \frac{1}{n}S'^2_{\overline{Y}} \qquad \text{não enviesados e consistentes}$$

$$\hat{V}(\hat{T}_{Bi}) = \frac{K^2}{n(n-1)}\sum_{i=1}^{n}(\overline{Y}_i-\overline{\overline{Y}})^2 = \frac{K^2}{n}S'^2_{\overline{Y}}$$

6.2.1.2 Amostra aleatória simples de UP

Igual probabilidade de selecção das UP: $\psi_i = \frac{1}{N}$ *i = 1, 2, ..., N*

A. Estimadores não enviesados:

$$\hat{\mu}_{Bi} = \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{K}\overline{Y}_i = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{M_o}\overline{Y}_i \quad \text{onde} \quad M_o = \frac{K}{N} \ \text{dimensão média das UP}$$

$$\hat{T}_{Bi} = \frac{N}{n}\sum_{i=1}^{n}M_i\overline{Y}_i = N\ \frac{\sum_{i=1}^{n}\hat{T}_i}{n} = N\overline{\hat{T}}$$

Estimadores para as variâncias:

$$\hat{V}(\hat{\mu}_{Bi}) = \frac{1}{n(n-1)}\ \sum_{i=1}^{n}\left(\frac{M_i}{M_o}\overline{Y}_i - \hat{\mu}_{Bi}\right)^2 \quad \text{não enviesados e consistentes}$$

$$\hat{V}(\hat{T}_{Bi}) = \frac{1}{n(n-1)}\ \sum_{i=1}^{n}\left(NM_i\overline{Y}_i - \hat{T}_{Bi}\right)^2 = \frac{1}{n(n-1)}\ \sum_{i=1}^{n}\left(N\hat{T}_i - \hat{T}_{Bi}\right)^2$$

B. Estimador de um rácio

Quando M_o (ou *K)* não é conhecido ter-se-á de estimar o seu valor. Tal estimativa terá por base as dimensões das UP que compõem a amostra

$$\hat{M}_o = \overline{M} = \frac{\sum_{i=1}^{n}M_i}{n}$$

Estimador

$$\hat{\mu}_{Bir} = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i\overline{Y}_i}{\overline{M}} = \frac{\sum_{i=1}^{n}M_i\overline{Y}_i}{\sum_{i=1}^{n}M_i} \qquad = \frac{\sum_{i=1}^{n}M_i\overline{Y}_i/n}{\sum_{i=1}^{n}M_i/n} = \frac{\overline{Z}}{\overline{M}}$$

Estimador do EQM

$$E\hat{Q}M(\hat{\mu}_{Bir}) \approx \frac{1}{\overline{M}^2 n}\left(S_Z'^2 + \hat{\mu}_{Bi}^2 S_M'^2 - 2\hat{\mu}_{Bi} S'_{ZM}\right)$$

onde

$$S_z'^2 = \frac{\sum_{i=1}^{n}(M_i\overline{Y}_i - \overline{Z})^2}{n-1} \quad , \quad S_M'^2 = \frac{\sum_{i=1}^{n}(M_i - \overline{M})^2}{n-1} \text{ e}$$

$$S'_{ZM} = \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})(M_i - \overline{M})}{n-1}$$

6.2.2. Selecção PISR nas duas etapas

Neste caso cada uma das unidades finais, u_{ij}, tem uma probabilidade de pertencer à amostra:

$P(j\text{-ésimo elemento da i-ésima UP ser seleccionado}) = \frac{n}{N} \times \frac{m_i}{M_i}$, *e* o seu peso será: $\omega_{ij} = \frac{N}{n} \times \frac{M_i}{m_i}$

A. Estimadores não enviesados:

Estimadores

$\hat{\mu}_{Bi} = \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{K}\overline{Y}_i = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{M_o}\overline{Y}_i$ onde $M_o = \frac{K}{N}$ dimensão média das UP

$\hat{T}_{Bi} = \frac{N}{n}\sum_{i=1}^{n}M_i\overline{Y}_i = N\frac{\sum_{i=1}^{n}\hat{T}_i}{n} = N\overline{\hat{T}}$ não enviesados

Variâncias dos estimadores

$$V(\hat{\mu}_{Bi}) = (1-\frac{n}{N})\frac{\sigma_{\overline{y}}'^2}{n} + \frac{N}{n}\sum_{i=1}^{N}(1-\frac{m_i}{M_i})\frac{M_i^2}{K^2}\frac{\sigma_i'^2}{m_i} =$$

$$= (1-f_1)\frac{\sigma_{\overline{y}}'^2}{n} + \frac{1}{Nn}\sum_{i=1}^{N}(1-f_{2i})\frac{M_i^2}{M_o^2}\frac{\sigma_i'^2}{m_i} = \frac{V(\hat{T}_{Bi})}{K^2}$$

onde, $\sigma_{\overline{y}}'^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\mu_i - \mu)^2$ e $\sigma_i'^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \mu_i)^2$

$$V(\hat{T}_{Bi}) = N^2(1-\frac{n}{N})\frac{\sigma_t'^2}{n} + \frac{N}{n}\sum_{i=1}^{N}(1-\frac{m_i}{M_i})M_i^2\frac{\sigma_i'^2}{m_i} =$$

$$= N^2(1-f_1)\frac{\sigma_t'^2}{n} + \frac{N}{n}\sum_{i=1}^{N}(1-f_{2i})M_i^2\frac{\sigma_i'^2}{m_i} = K^2V(\hat{\mu}_{Bi})$$

onde, $\sigma_t'^2 = \frac{1}{N-1}\sum_{i=1}^{N}(t_i - \mu_t)^2$ e $\mu_t = \frac{t_y}{N}$

Estimadores para as variâncias:

$$\hat{V}(\hat{\mu}_{Bi}) = (1-f_1)\frac{S_e'^2}{n} + \frac{1}{Nn}\sum_{i=1}^{n}(1-f_{2i})\frac{M_i^2}{M_o^2}\frac{S_i'^2}{m_i} = \frac{\hat{V}(\hat{T}_{Bi})}{K^2}$$ Estimadores consistentes

variação ***entre*** UP + variação ***dentro*** das UP

onde,

$$S_e'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{M_i}{M_o}\overline{Y}_i - \hat{\mu}_{Bi})^2 \text{ e } S_i'^2 = \frac{1}{m_i-1}\sum_{j=1}^{m_i}(Y_{ij} - \overline{Y}_i)^2$$

$$\hat{V}(\hat{T}_{Bi}) = N^2(1-f_1)\frac{S_t'^2}{n} + \frac{N}{n}\sum_{i=1}^{n}(1-f_{2i})M_i^2\frac{S_i'^2}{m_i} = K^2\hat{V}(\hat{\mu}_{Bi})$$

variação ***entre*** UP + variação ***dentro*** das UP

onde,

$$S_t'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{T}_i - \overline{\hat{T}})^2 \quad onde \quad \overline{\hat{T}} = \frac{\hat{T}_{Bi}}{N}$$

B. Estimador de um rácio

Quando M_o (ou K) não é conhecido ter-se-á de estimar o seu valor.

Tal estimativa terá por base as dimensões das UP que compõem a amostra

$$\hat{M}_o = \overline{M} = \frac{\sum_{i=1}^{n} M_i}{n}$$

Estimador e seu EQM

$$\hat{\mu}_{Bir} = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i\overline{Y}_i}{\overline{M}} = \frac{\sum_{i=1}^{n} M_i\overline{Y}_i}{\sum_{i=1}^{n} M_i} = \frac{\sum_{i=1}^{n} M_i\overline{Y}_i / n}{\sum_{i=1}^{n} M_i / n} = \frac{\overline{Z}}{\overline{M}}$$

$$E\hat{Q}M(\hat{\mu}_{Bi_r}) \approx \frac{1}{\overline{M}^2}\left[(1-f_1)\frac{S_r'^2}{n} + \frac{1}{nN}\sum_{i=1}^{n}M_i^2(1-f_{2i})\frac{S_i'^2}{m_i}\right]$$

onde $S_r'^2 = \dfrac{\sum_{i=1}^{n}(M_i\overline{Y}_i - M_i\hat{\mu}_{Bi_r})^2}{n-1}$ e $S_i^2 = \dfrac{1}{m_i-1}\sum_{j=1}^{m_i}(Y_{ij} - \overline{Y}_i)^2$

A opção pela escolha das UP com probabilidades iguais é tomada muitas vezes quando:

1. As UP têm dimensões semelhantes. Neste caso $M_1 \approx M_2 \approx ... \approx M_N = M$ e a dimensão das amostras seleccionadas em cada UP (2ª etapa) também devem ser iguais, pois, assim

cada uma das unidades finais, u_{ij}, tem a mesma probabilidade de pertencer á amostra (todos os elementos têm o mesmo peso).
Assim, $m_1 = m_2 = ... = m_n$ e $f_{2i} = f_2 \quad para \quad i = 1,2,...,n$

$$\hat{\mu}_{Bi} = \frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i = \overline{\overline{Y}}$$

$$\hat{V}(\hat{\mu}_{Bi}) = = \frac{(1-f_1)}{n(n-1)}\sum_{i=1}^{n}(\overline{Y}_i - \overline{\overline{Y}})^2 + \frac{1-f_2}{Nn}\sum_{i=1}^{n}\frac{S_i^{'2}}{m}$$

onde, $S_i^{'2} = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(Y_{ij} - \overline{Y}_i)^2$

2. Não existe informação adicional sobre o universo e portanto, não se conhecem as dimensões das UP.
   Neste caso habitualmente dimensiona-se a amostra da 2ª etapa de forma a que as taxas de amostragem sejam constantes, fazendo com que a amostra seja autoponderada (todos os elementos têm o mesmo peso).

$$f_{2i} = \frac{m_i}{M_i} = f_2 \Rightarrow m_i = f_2 M_i \quad para \quad i = 1,2,...,n$$

6.3. Determinação da dimensão da amostra

Princípio geral

Qual o nº de unidades primárias e secundárias a seleccionar para garantir, com uma confiança de ***100(1-α) %***, um nível de precisão ***e*** para as estimativas a obter?

Dois níveis

⇒ qual o número ***n*** de *unidades primárias* a seleccionar na primeira etapa de amostragem

⇒ quais os valores para $\boldsymbol{m_i}$, ou seja, quantas *unidades secundárias* deverão ser seleccionadas em cada uma das unidades primárias escolhidas

Assim,

$$Z = \frac{\hat{\mu}_{Bi} - \mu}{\sqrt{V(\hat{\mu}_{Bi})}} \overset{a}{\sim} N(0;1)$$

$$n, m_i : P(|\hat{\mu}_{Bi} - \mu| < e) = 1 - \alpha \Leftrightarrow n, m_i : P(\frac{|\hat{\mu}_{Bi} - \mu|}{\sqrt{V(\hat{\mu}_{Bi})}} < \frac{e}{\sqrt{V(\hat{\mu}_{Bi})}}) = 1 - \alpha$$

$$n, m_i : \quad V(\hat{\mu}_{Bi}) = \frac{e^2}{z^2_{\alpha/2}} \quad \textbf{[1]}$$

onde $V(\hat{\mu}_{Bi})$ é função de ***n*** e dos $\boldsymbol{m_i}$

6.3.1. Probabilidade de selecção das UP proporcional à dimensão (pps)

Tratando, por exemplo, o caso em que as unidades primárias são seleccionadas com probabilidades proporcionais à sua dimensão, tem-se que

$$\sigma_n^2 = V(\overline{\overline{Y}}) = \frac{1}{n}\left[\sum_{i=1}^{N}\frac{M_i}{K}\frac{\sigma_i^2}{m_i} + \sigma_e^2\right] \quad \text{onde } \sigma_e^2 = \sum_{i=1}^{m}\frac{M_i}{K}(\mu_i - \mu)^2$$

Substituindo em **[1]**

$$\frac{1}{n}\left[\sum_{i=1}^{N}\frac{M_i}{K}\frac{\sigma_i^2}{m_i} + \sigma_e^2\right] = \frac{\delta^2}{u_\alpha^2}$$

Esta expressão evidencia que, fixado o número de unidades primárias a seleccionar na 1ª etapa, o dimensionamento da amostra apenas à custa do número de unidades secundárias, m_i, **pode não permitir** atingir os níveis de precisão desejados (pois a 2ª parcela não depende de m_i), ou seja:

a variável fundamental para garantir os níveis de precisão fixados é o número de unidades primárias (n) seleccionadas na 1ª etapa (não os m_i),

Resolvendo em ordem a n obtém-se:

$$n = \frac{\left[\sum_{i=1}^{N}\frac{M_i}{K}V(\overline{Y}_i) + \sigma_e^2\right] \times z_{\alpha/2}^2}{e^2} \qquad \textbf{[2]}$$

e esta igualdade só será operacional se forem conhecidos os valores de $\sum_{i=1}^{N}\frac{M_i}{K}V(\overline{Y}_i)$ e de σ_e^2, ou seja as variâncias dentro das unidades primárias e a variância das médias dessas unidades na população em causa.

Na prática esta questão é resolvida tentando obter estimativas desses parâmetros por um destes processos:

- recorrendo ao conhecimento que possa haver de tais parâmetros, numa população que possa considerar-se semelhante e, como precaução contra um eventual sub-dimensionamento da amostra, poder-se-á dar uma certa margem de segurança, trabalhando com uma dimensão maior do que a calculada com esses parâmetros;
- obtendo uma **pré-amostra** com n_o **unidades primárias na 1ªetapa, seleccionadas com probabilidade proporcional à dimensão e utilizando o estimador**

$$S'^2_{\overline{\overline{Y}}} = n_o\hat{V}(\hat{\mu}_{Bi}) = \frac{1}{(n_o\text{-}1)}\sum_{i=1}^{n_o}(\overline{Y}_i - \overline{\overline{Y}})^2 \qquad \textbf{[3]}$$

que é um **estimador não enviesado e consistente** de $\left[\sum_{i=1}^{N}\frac{M_i}{K}V(\overline{Y}_i) + \sigma_e^2\right]$

Assim, fixa-se o mínimo de unidades finais a seleccionar dentro de cada UP ($b = min\{m_i\}$), substitui-se na expressão **[2]** $\left[\sum_{i=1}^{N}\frac{M_i}{K}V(\bar{Y}_i)+\sigma_e^2\right]$ pela sua estimativa, obtendo-se o número de unidades primárias na amostra (1ªetapa),

$$n=\frac{z^2_{\alpha/2}\times S'^2_{\bar{Y}}}{e^2} \qquad \text{com} \quad S'^2_{\bar{Y}}=\frac{1}{(n_o\text{-}1)}\sum_{i=1}^{n_o}(\bar{Y}_i-\bar{\bar{Y}})^2$$

variância corrigida das médias amostrais das UP.

6.3.2. Selecção PISR nas duas etapas

$$n, m_i: \quad V(\hat{\mu}_{Bi})=\frac{e^2}{z^2_{\alpha/2}} \qquad \textbf{[1]}$$

onde $V(\hat{\mu}_{Bi})$ é função de ***n*** e dos m_i .

$$V(\hat{\mu}_{Bi})=(1-\frac{n}{N})\frac{\sigma'^2_{\bar{y}}}{n}+\frac{N}{n}\sum_{i=1}^{N}(1-\frac{m_i}{M_i})\frac{M_i^2}{K^2}\frac{\sigma'^2_i}{m_i}=$$

$$=\frac{\sigma'^2_{\bar{y}}}{n}-\frac{\sigma'^2_{\bar{y}}}{N}+\frac{1}{Nn}\sum_{i=1}^{N}(1-f_{2i})\frac{M_i^2}{M_o^2}\frac{\sigma'^2_i}{m_i}$$

onde, $\sigma'^2_{\bar{y}}=\frac{1}{N-1}\sum_{i=1}^{N}(\mu_i-\mu)^2$ e $\sigma'^2_i=\frac{1}{M_i-1}\sum_{j=1}^{M_i}(y_{ij}-\mu_i)^2$

Substituindo em [1]

$$\frac{\sigma'^2_{\bar{y}}}{n}-\frac{\sigma'^2_{\bar{y}}}{N}+\frac{1}{Nn}\sum_{i=1}^{N}(1-f_{2i})\frac{M_i^2}{M_o^2}\frac{\sigma'^2_i}{m_i}=\frac{e^2}{z^2_{\alpha/2}}$$

e resolvendo em ordem a ***n,***

$$n=\frac{z^2_{\alpha/2}(\sigma'^2_{\bar{Y}}+\frac{1}{N}\sum_{i=1}^{N}(1-f_{2i})\frac{M_i^2}{M_o^2}\frac{\sigma'^2_i}{m_i})}{e^2+\frac{\sigma'^2_{\bar{Y}}z^2_{\alpha/2}}{N}}$$

Utilizando estimativas obtidas em sondagens anteriores ou na pré-amostra

$$n=\frac{z^2_{\alpha/2}\left[S'^2_e+\frac{1}{N}\sum_{i=1}^{n_o}(1-f_{2i})\frac{M_i^2}{M_o^2}\frac{S'^2_i}{m_i}\right]}{e^2+\frac{S'^2_e z^2_{\alpha/2}}{N}}$$

7. NÃO RESPOSTA

Leitura obrigatória: capítulo 8 do livro "Sampling: Design and Analysis", Sharon L. Lohr

7.1. Introdução

O que é a não resposta ? (nonresponse ou missing data)

- falha na obtenção de informação nalgum dos momentos do processo de recolha de dados.
- ocorre durante a recolha propriamente dita, depois da amostra ser seleccionada.
- pode provocar enviesamentos na estimação dos parâmetros da população susceptíveis de afectar a validade dos resultados.

Calcular taxas de não resposta.

Factores que influenciam a "não resposta"

⇒ Objecto da sondagem (questões delicadas)
⇒ Momento da inquirição (férias/festividades ou momentos de grande actividade/trabalho...)
⇒ Método de recolha da informação (inquérito postal, telefónico, directo...)
⇒ Entrevistadores
⇒ Instrumento de notação (Questionário: relevância e clareza das questões, extensão
⇒ Introdução do inquérito: explicação do porquê do inquérito , importância
⇒ Incentivos à resposta / penalizações para a não resposta; Avisos preliminares pedindo colaboração, insistências, seguimento …

Qualidade dos dados recolhidos altamente determinada pela fase de planeamento do trabalho: **PREVENIR ! ! ! … PREVENIR ! ! ! …**

Tipos de "não resposta"

❖ ***Não resposta total***: falta toda a informação sobre a UA ("nr").

- UA indevidamente na base de sondagem (inelegíveis): erros na base de sondagem utilizada (desactualizadas, incompletas, duplas contagens…) – ***amostras de substituição ou regras claras de substituição***;
- UA não encontrada no local previsto (ausências mais ou menos prolongadas…) – ***novos contactos em horários diferentes***;
- UA localizada mas não participa: a unidade estatística existe, foi contactada, mas não participa (quer por negligência quer por recusa) – ***tentar modelar o comportamento desta fracção, incentivos à obtenção de alguma informação, insistências e dupla amostragem, pós-estratificação…***

❖ ***Não resposta parcial***: falta de resposta a alguns itens do questionário

- falta de informação sobre alguma(s) variável(is) ou erros de resposta a alguns itens do questionário, não corrigidos na totalidade, ou mesmo recusa de responder a certas questões – ***imputação e ajustamentos amostrais.***

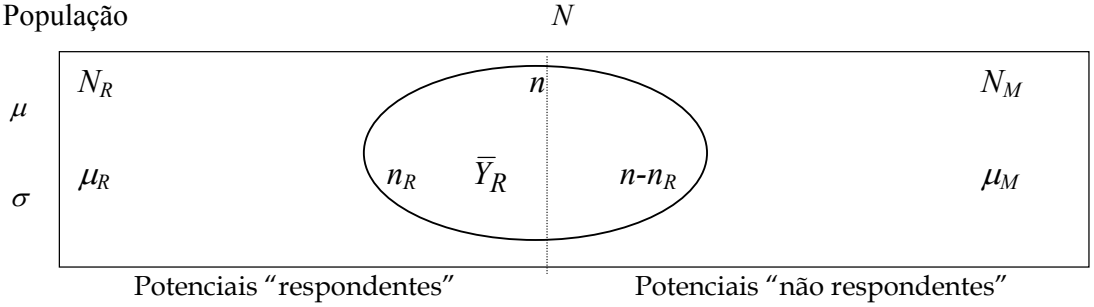
7.1.1. Abordagens ao problema

- Ignorar a "não resposta" não é recomendável mas é muitas vezes a situação mais comum.
- **PREVENIR**: tentar minorar o seu aparecimento preparando cuidadosamente os vários aspectos da sondagem (que é melhor método – mais vale prevenir do que remediar…);
- Seleccionar uma subamostra representativa dos que não responderam; usá-la para inferir sobre os "não respondentes";
- Utilizar um modelo para "prever" os valores das variáveis para os "não respondentes" através de metodologias de ajustamentos amostrais/pós-estratificação ou de imputação.

7.1.2. Consequências teóricas

O principal problema causado pela não resposta é o ***potencial enviesamento*** das estimativas para a totalidade da população, calculadas somente com base nos que responderam.

Pense-se então na população como estando dividida (embora de forma artificial) em dois estratos: "respondentes" e "não respondentes".



onde $\mu = \frac{N_R}{N}\mu_R + \frac{N_M}{N}\mu_M$.

Seleccionou-se uma amostra aleatória simples de dimensão n onde se obtiveram n_R respostas correspondendo a uma média de $\overline{Y}_R$.

Mas, certamente que nestas respostas:

→ $\overline{Y}_R$ é um estimador não enviesado de μ_R, ou seja, $E(\overline{Y}_R) = \mu_R$.

→ No entanto ao considerar $\overline{Y}_R$ como estimador de μ tal estimador é enviesado, sendo o enviesamento dado por:

$$Env(\overline{Y}_R) = \mu_R - \mu = \mu_R - \frac{N_R}{N}\mu_R - \frac{N_M}{N}\mu_M = \frac{N_M}{N}(\mu_R - \mu_M).$$

O enviesamento depende da taxa de não resposta ($\frac{N_M}{N}$) e da diferença entre o valor médio para os dois grupos os que respondem e os que não respondem $(\mu_R - \mu_M)$.

O EQM do estimador será assim:

$$EQM(\overline{Y}_R) = V(\overline{Y}_R) + \left[Env(\overline{Y}_R)\right]^2 = (1 - \frac{n}{N})\frac{\sigma_R'^2}{n_R} + \frac{N_M^2}{N^2}(\mu_R - \mu_M)^2$$

7.1.3. Mecanismos que originam a "não resposta"

Seja uma v.a. definida por: $R_i = \begin{cases} 1 & se\ U_i\ responde \\ 0 & caso\ contrário \end{cases}$

Defina-se como ϕ_i a probabilidade de U_i responder (propensão á resposta), desconhecida mas diferente de zero: $\phi_i = Pr\{R_i = 1\}$

Designe por: y_i a variável inquirida na sondagem e x_i o vector de informações conhecidas sobre U_i onde se inclui a informação que serviu de base ao desenho da amostra.

- **Não resposta completamente aleatória** (MCAR - missing completely at random)
  - A propensão para responder é completamente aleatória, não está relacionada nem com y nem com x e, assim, os "respondentes" são representativos da amostra seleccionada. Isto implica que:
    - Seleccionada uma amostra aleatória simples de n unidades estatísticas os "respondentes" correspondem a uma amostra aleatória simples de dimensão n_R.
    - $\overline{Y}_R$ (média amostral das respostas) é um estimador não enviesado para a média da população μ.
  - Este mecanismo está implicitamente assumido quando se ignora a não resposta (muitas vezes erradamente).

- **Não resposta aleatória dado x** (MAR - missing at random given covariates)
  - A propensão para responder depende de x, características conhecidas para toda a amostra, mas não está relacionada com y. Isto implica que uma vez conhecidos os valores de x, podem proceder-se a ajustamentos amostrais que tenham em conta a não resposta, eliminando assim o seu efeito.
  - Por exemplo, a propensão para responder pode depender somente do género e da idade da pessoa inquirida, mas não estar relacionada com as variáveis em estudo.

- **Não resposta não-ignorável**
  - A propensão para responder depende da variável de interesse (y) e não pode ser completamente explicada pelos outros factores (x).
  - A modelação poderá não eliminar completamente o potencial enviesamento provocado pela não resposta.

7.2. Pesquisa de factores explicativos

Descrição por análise de dados

Estabelecer, mesmo que grosseiramente, tipologias de não respondentes, de modo a corrigir os enviesamentos que possam ocorrer na interpretação dos resultados e a caracterizar a não resposta para compreendê-la e minorá-la em futuros estudos.

Podem utilizar-se os instrumentos clássicos de análise estatística multivariada, nomeadamente: análise de componentes principais e análise discriminante, para realçar

variáveis caracterizadoras conhecidas e relacioná-las com variáveis ligadas às reacções face aos inquéritos (recusa ou não, tipo de não resposta…)

Modelos econométricos

Conhecendo factores susceptíveis de influenciar o comportamento "resposta/não resposta", pode-se tentar modelar estas influências utilizando métodos econométricos para variáveis qualitativas.

Variável endógena

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

Variáveis explicativas

$X_1, X_2, ..., X_p$, supostas conhecidas

A probabilidade de resposta (ou de não resposta) é função das variáveis $X_1, X_2, ..., X_p$. Assim, $P(R_i = 1) = F(X_{1i}, X_{2i}, ..., X_{pi}) = F_i$.

A verosimilhança da amostra vem dada por: $L = \prod_{i=1}^{n} F_i^{r_i} (1 - F_i)^{1-r_i}$, estimando-se os parâmetros da função *F* pelo método da máxima verosimilhança.

Os modelos distinguem-se pela escolha da função *F*. Os mais utilizados são os modelos Logit e Probit.

a) Modelo Logit : $F_i = \dfrac{1}{1 + exp\left\{-(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ji}\right\}}$ (logística)

Neste modelo se βj>0 um acréscimo de Xji traduz-se num aumento de Fi, se βj<0 um acréscimo de Xji traduz-se numa diminuição de Fi.

b) Modelo Probit: $F_i = \Phi(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ji})$ onde Φ é a função de distribuição da normal estandardizada. Os parâmetros β_j têm interpretação semelhante aos do modelo anterior.

7.3. Metodologias de tratamento

7.3.1. Métodos de ajustamento amostral

Estes métodos tratam o problema da não resposta ajustando os dados obtidos das respostas de modo a serem utilizados para a estimação, tentando compensar a falta de resposta.

7.3.1.1 Caso geral

Genericamente sabe-se que um estimador para o total pode ser dado por: $\hat{T}_y = \sum_{i \in S} \omega_i Y_i$, onde ω_i, peso amostral de cada elemento da amostra, é o inverso da probabilidade de selecção ($\omega_i = 1/\pi_i$), mas isso pressupõe que não existem problemas de "não resposta". Por exemplo, quando PISR $\omega_i = N/n$, representa o número de elementos da população que estão representados por este elemento da amostra. No caso de existirem não respostas estes pesos devem ser ajustados de forma a ter em conta essa não resposta, ou seja calcular a probabilidade da i-ésima unidade pertencer à amostra e responder = $\pi_i \phi_i$ onde os ϕ_i,

propensão à resposta, têm de ser estimadas para cada uma das unidades finais, com base em variáveis conhecidas para a população. Assim, o peso ajustado para cada uma das unidades *que responderam* $\tilde{\omega}_i = \frac{\omega_i}{\hat{\phi}_i} = \frac{1}{\pi_i \hat{\phi}_i}$, sendo os estimadores calculados com base nas respostas:

$$\hat{T}_y = \sum_{i \in Resp} \tilde{\omega}_i Y_i \text{ e} \qquad \hat{\mu}_y = \frac{\hat{T}_y}{\sum_{i \in Resp} \tilde{\omega}_i}.$$

EXEMPLO:

Uma amostra de 1000 indivíduos retirada de uma população com 150104, em que cada indivíduo foi seleccionado com uma probabilidade π_i conhecida (logo, tb se conhecem os pesos $\omega_i = 1/\pi_i$), por exemplo, no caso de selecção PISR $\pi_i = 1000/150104 = 0{,}0067$ e $\omega_i = 1/\pi_i = 150{,}104$.

Suponha que se conhece a classe etária de cada um dos indivíduos seleccionados, então poder-se-á utilizar as classes etárias como classes de ajustamento para tratamento da não resposta.

	Classes etárias					
	15-24	25-34	35-44	45-64	65 e +	Total
Repartição da amostra	202	220	180	195	203	1000
Repartição das respostas	124	187	162	187	203	863
$\sum_{i \in S} \omega_{ic}$ na amostra	30322	33013	27046	29272	30451	150104
ω_{ic} (média na amostra)	150,1089	150,0591	150,2556	150,1128	150,0049	
$\sum_{i \in Resp} \omega_{ic}$ na resposta	18693	28143	24371	28138	30451	
$\hat{\phi}_{ic} = \sum_{i \in Resp} \omega_{ic} \Big/ \sum_{i \in S} \omega_{ic}$	0,6165	0,8525	0,9011	0,9613	1,0000	
Factor de ajustamento	1,622	1,173	1,110	1,040	1,000	
$\tilde{\omega}_{ic} = \omega_{ic} / \hat{\phi}_{ic}$ (média)	243,4923	176,026	166,7478	156,1626	150,0049	

(exemplo do livro)

7.3.1.2 Ajustamento por estratificação a posteriori

Os dados da amostra observada (n_R) são estratificados a posteriori, segundo critérios de que se conheça a repartição exacta na população:

$$W_h = \frac{N_h}{N} \quad h = 1,2,...,H.$$

Por exemplo, para inquéritos junto às famílias podem utilizar-se variáveis sócio-demográficas: nº de pessoas do agregado, idade do chefe de famílias, CSP, categoria do município, …

A partir desta pós-estratificação, calcula-se para cada estrato um coeficiente permitindo restabelecer a parte que representa no total da população.

Substituem-se, assim, as não respostas da amostra pelos valores médios das respostas obtidas nos estratos criados a posteriori.

Este método tem como hipótese subjacente que: para cada estrato a população de "não respondentes" tem comportamento idêntico à dos respondentes, isto só se verifica se no interior de cada estrato a probabilidade de não resposta não depender da variável em estudo.

Amostra observada	$\longrightarrow$	n_R	$\overline{Y}_R$	S'^2_R

Estratos	1 .	2 …	h …	H.	
A posteriori	n_{1R}	n_{2R}	n_{hR}	n_{HR}	Dimensão
	$\overline{Y}_{1R}$	$\overline{Y}_{2R}$	$\overline{Y}_{hR}$	$\overline{Y}_{HR}$	Média
	S'^2_{1R}	S'^2_{2R}	S'^2_{hR}	S'^2_{HR}	Variância corrigida

O estimador não enviesado para μ: $\hat{\mu}_{Post} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{Y}_{hR} = \sum_{h=1}^{H} W_h \overline{Y}_{hR}$

que tem variância estimada aproximadamente igual a:

$$\hat{V}(\hat{\mu}_{post}) \approx (1-\frac{n_R}{N})\frac{1}{n_R}\sum_{h=1}^{H} W_h S'^2_{hR} + (1-\frac{n_R}{N})\frac{1}{n_R^2}\sum_{h=1}^{H}(1-W_h)S'^2_{hR}$$

$$\boxed{\approx (1-\frac{n_R}{N})\frac{1}{n_R}\sum_{h=1}^{H} W_h S'^2_{hR}}$$

ou, alternativamente, também se pode utilizar a aproximação

$$\hat{V}(\hat{\mu}_{post}) \approx \sum_{h=1}^{H} W_h^2 (1-\frac{n_{hR}}{N_h})\frac{1}{n_{hR}} S'^2_{hR}$$

7.3.2. Insistência e amostragem em duas fases

(Subamostragem das não respostas- Dupla amostragem).

A ideia de base é a de que, em certas condições, pode obter-se informações relativas à subpopulação (estrato) dos "não respondentes".

Hansen e Hurwitz (1946) propuseram inquirir uma subamostra de "não respondentes" e usar as metodologias de estratificação por dupla amostragem para estimar a média ou total da população. Selecciona-se uma amostra aleatória simples de dimensão n: n_R respondem e n_M não respondem.

	População			Amostra		
				1ª fase		2ª fase
Respondem	N_R	μ_R	σ'_R	n_R	$\bar{Y}_R, S'^2_R$	
Não respondem	N_M	μ_M	σ'_M	n_M	→	$n'_M, \bar{Y}_M, S'^2_M$
TOTAL	N	μ	σ'	n		

De entre os n_M que não responderam selecciona-se uma amostra aleatória simples de dimensão $n'_M = \nu \times n_M$, onde 0<ν<1 é a taxa de insistência, e recolhe-se a informação para esses elementos (2ª fase de amostragem).

Os estimadores para a média e o total serão dados por:

$$\hat{\mu} = \frac{n_R}{n}\bar{Y}_R + \frac{n_M}{n}\bar{Y}_M \quad \text{e} \quad \hat{T}_y = N\hat{\mu} = \frac{N}{n}(n_R\bar{Y}_R + n_M\bar{Y}_M) = \frac{N}{n}\sum_{i=1}^{n_R} Y_i + \frac{N}{n\nu}\sum_{i=1}^{n'_M} Y_i$$

com variância do estimador para a média igual a:

$$V(\hat{\mu}) = (1-\frac{n}{N})\frac{\sigma'^2}{n} + \frac{1}{n}\frac{N_M}{N}(\frac{1-\nu}{\nu})\sigma'^2_M$$

$$= \frac{N-n}{Nn}\left[\frac{N_R N_M}{N^2}(\mu_R - \mu_M)^2 + \frac{N_R}{N}\sigma'^2_R + \frac{N_M}{N}\sigma'^2_M\right] + \frac{1}{n}\frac{N_M}{N}(\frac{1-\nu}{\nu})\sigma'^2_M$$

sendo esta variância estimada por:

$$\hat{V}(\hat{\mu}) = \frac{N-n}{Nn}\left[\frac{n_R n_M}{n^2}(\bar{Y}_R - \bar{Y}_M)^2 + \frac{n_R}{n}S'^2_R + \frac{n_M}{n}S'^2_M\right] + \frac{1}{n}\frac{n_M}{n}(\frac{1-\nu}{\nu})S'^2_M$$

Se todos os elementos seleccionados para a subamostra responderem (amostra da 2ª fase), esta metodologia não só elimina o enviesamento provocado pela não resposta original como também a tem em conta na estimação da variância.

7.3.2.1 Taxa de amostragem óptima entre as não respostas

Seleccionar uma subamostra dos que não responderam e fazer um esforço intensivo para recolher informação sobre esses elementos: $n'_M = \nu \times n_M$.

Os valores de n e ν são calculados de forma a proporcionar uma determinada precisão pelo menor custo possível.

Seja o custo de selecção da amostra global igual a:

$$C = c_o \times n + c_1 \times n_R + c_2 \times n'_M = c_o \times n + c_1 \times n_R + c_2 \times n_M \times \nu \quad \text{onde,}$$

c_o – custo unitário da realização da 1ª recolha (1ª fase)

c_1 – custo unitário de carregamento e processamento das respostas à 1ª fase

c_2 – custo unitário de obtenção, carregamento e processamento das respostas obtidas nas insistências (2ª fase).

O valor esperado do custo será:

$$E(C) = c_o \times n + c_1 \times \frac{N_R}{N} n + c_2 \times \frac{N_M}{N} n \times \nu = n\left(c_o + c_1 \times \frac{N_R}{N} + c_2 \times \frac{N_M}{N} \times \nu\right)$$

De um modo geral, para este dimensionamento, o problema a resolver será um destes dois

I ou ***II***

$$n,\nu : Min\ E(C)$$
$$s.a.$$
$$V(\hat{\mu}) \le V_o$$

$$n,\nu : Min\ V(\hat{\mu})$$
$$s.a.$$
$$E(C) \le C$$

<u>Problema I</u>

Calcular a dimensão inicial n e a taxa de insistência ν (dimensão da subamostra) de modo a minimizar o custo médio, sujeito à restrição de que a variância do estimador não seja superior a V_o ,ou seja, satisfazendo uma determinada precisão :

$$n,v,v:M\left[n\left(c_o + c_1 \times \frac{N_R}{N} + c_2 \times \frac{N_M}{N} \times v\right)\right]$$

$$s.a.$$

$$(1-\frac{n}{N})\frac{\sigma'^2}{n} + \frac{1}{n}\frac{N_M}{N}(\frac{1-v}{v})\sigma'^2_M \le V_o$$

Sendo o custo médio uma função crescente de ***n*** então convém escolher para valor de ***n*** o menor possível satisfazendo a restrição de precisão, isto é:

$$n: \frac{\sigma'^2}{n} - \frac{\sigma'^2}{N} + \frac{1}{n}\frac{N_M}{N}(\frac{1-\nu}{\nu})\sigma'^2_M = V_o \Rightarrow$$

$$n_{opt} = \frac{N\sigma'^2 + N_M(\frac{1-\nu}{\nu})\sigma'^2_M}{NV_o + \sigma'^2} = \frac{\sigma'^2 + \frac{N_M}{N}(\frac{1-\nu}{\nu})\sigma'^2_M}{V_o + \frac{\sigma'^2}{N}}$$

A determinação da taxa óptima de insistência, ν_{opt}, efectua-se minimizando

$$\nu : Min\left[\frac{N\sigma'^2 + N_M(\frac{1-\nu}{\nu})\sigma'^2_M}{NV_o + \sigma'^2}\left(c_o + c_1 \times \frac{N_R}{N} + c_2 \times \frac{N_M}{N} \times \nu\right)\right]$$

Resolvendo obtém-se:

$$\nu_{opt} = \sqrt{\frac{\sigma_M'^2 (c_o + \frac{N_R}{N} c_1)}{c_2 (\sigma'^2 - \frac{N_M}{N} \sigma_M'^2)}} \qquad \frac{N_R}{N} \quad \textit{taxa de resposta} \qquad \frac{N_M}{N} \quad \textit{taxa de não resposta}$$

Problema II

Calcular a dimensão inicial n e a taxa de insistência ν (dimensão da subamostra) de modo a minimizar a variância do estimador mas satisfazendo uma restrição orçamental:

$$\boldsymbol{n, \nu : Min} \; (1 - \frac{\boldsymbol{n}}{\boldsymbol{N}}) \frac{\boldsymbol{\sigma}'^2}{\boldsymbol{n}} + \frac{1}{\boldsymbol{n}} \frac{\boldsymbol{N_M}}{\boldsymbol{N}} (\frac{1 - \boldsymbol{\nu}}{\boldsymbol{\nu}}) \boldsymbol{\sigma}_{\boldsymbol{M}}'^2$$

$$\textbf{s.a.}$$

$$\boldsymbol{n} \left(\boldsymbol{c_o} + \boldsymbol{c}_1 \times \frac{\boldsymbol{N_R}}{\boldsymbol{N}} + \boldsymbol{c}_2 \times \frac{\boldsymbol{N_M}}{\boldsymbol{N}} \times \boldsymbol{\nu} \right) \leq \boldsymbol{C}$$

Sendo a variância do estimador uma função decrescente de $\boldsymbol{n}$ então convém escolher para valor de $\boldsymbol{n}$ o maior possível que satisfaça a restrição orçamental, isto é, tal que:

$$n : \boldsymbol{n} \left(\boldsymbol{c_o} + \boldsymbol{c}_1 \times \frac{\boldsymbol{N_R}}{\boldsymbol{N}} + \boldsymbol{c}_2 \times \frac{\boldsymbol{N_M}}{\boldsymbol{N}} \times \boldsymbol{\nu} \right) = \boldsymbol{C}, \Rightarrow \quad \boldsymbol{n_{opt}} = \frac{\boldsymbol{NC}}{\boldsymbol{c_o N} + \boldsymbol{c}_1 \times \boldsymbol{N_R} + \boldsymbol{c}_2 \times \boldsymbol{N_M} \times \boldsymbol{\nu}}$$

A determinação da taxa óptima de insistência, ν_{opt}, efectua-se minimizando

$$\underset{\boldsymbol{\nu}}{\boldsymbol{Min}} \left[\boldsymbol{\sigma}'^2 \frac{\boldsymbol{c_o N} + \boldsymbol{c}_1 \times \boldsymbol{N_R} + \boldsymbol{c}_2 \times \boldsymbol{N_M} \times \boldsymbol{\nu}}{\boldsymbol{NC}} - \frac{\boldsymbol{\sigma}'^2}{\boldsymbol{N}} + \frac{\boldsymbol{c_o N} + \boldsymbol{c}_1 \times \boldsymbol{N_R} + \boldsymbol{c}_2 \times \boldsymbol{N_M} \times \boldsymbol{\nu}}{\boldsymbol{NC}} \frac{\boldsymbol{N_M}}{\boldsymbol{N}} (\frac{1 - \boldsymbol{\nu}}{\boldsymbol{\nu}}) \boldsymbol{\sigma}_{\boldsymbol{M}}'^2 \right]$$

Resolvendo obtém-se:

$$\boldsymbol{\nu_{opt}} = \sqrt{\frac{\boldsymbol{\sigma}_{\boldsymbol{M}}'^2 (\boldsymbol{c_o} + \frac{\boldsymbol{N_R}}{\boldsymbol{N}} \boldsymbol{c}_1)}{\boldsymbol{c}_2 (\boldsymbol{\sigma}'^2 - \frac{\boldsymbol{N_M}}{\boldsymbol{N}} \boldsymbol{\sigma}_{\boldsymbol{M}}'^2)}} \qquad \frac{\boldsymbol{N_R}}{\boldsymbol{N}} \quad \textbf{\textit{taxa de resposta}} \qquad \frac{\boldsymbol{N_M}}{\boldsymbol{N}} \quad \textbf{\textit{taxa de não resposta}}$$

7.3.3. Imputação de respostas – Estratégia na fase de estimação

Utilizado sobretudo para resolver o problema de não resposta parcial. Métodos para preencher valores de variáveis não respondidas tendo em conta informação auxiliar que permita imputar com alguma objectividade os valores em falta.

De qualquer modo dever-se-á sempre assinalar no ficheiro quando os valores resultem de imputação.

- ***Imputação dedutiva***
- ***Imputação média***
- ***Imputação média da classe***
- ***Imputação por regressão***
- ***Múltipla imputação***
- ***Imputação a quente (Hot-Deck)*** – imputação feita com base em dados observados no próprio inquérito.
  - ***Imputação aleatória***
  - ***Imputação sequencial***
  - ***Imputação através de uma função distância***
- ***Imputação a frio (Cold-Deck)*** – imputação feita com base em dados de inquéritos anteriores ou qq outra informação anteriormente recolhida.

7.3.4. Método das respostas aleatórias

OBJECTIVO: Limitar o número de não respostas ou mais geralmente a influência dos erros de medida devidos a *respostas voluntariamente erradas*.

→ Convencer o inquirido do ***anonimato da sua resposta*** sobretudo quando as questões são "melindrosas".

[Warner, 1965]

Seja por exemplo a questão A:

→ "Já alguma vez defraudou o fisco?" (SIM / NÃO)

Objectivo: Estimar a proporção de indivíduos que já defraudaram o fisco, P_A

Estratégia a utilizar:

- De um saco com θ bolas brancas e (1-θ) bolas pretas o inquirido retira ao acaso uma e vê a cor, sem mostrar o resultado ao inquiridor.

- Se a bola é *branca* responde SIM ou NÃO à afirmação " É verdade que já defraudei o fisco".
  Se a bola é *preta* responde SIM ou NÃO á afirmação " É verdade que nunca defraudei o fisco".

- O inquiridor regista a resposta do inquirido SIM ou NÃO ignorando a que afirmação corresponde.

Seja ϕ - a probabilidade de se obter uma resposta SIM no inquérito. Então,

ϕ **= P(Bola branca)P(Sim/Bola branca)+P(Bola preta) P(Sim/Bola preta) =**

$$= \theta \times P_A + (1-\theta) \times (1-P_A) = 1-\theta + P_A \times (2\theta - 1)$$

e $$P_A = \frac{\phi - (1-\theta)}{2\theta - 1} \quad \text{com} \quad \theta \neq 0{,}5$$

E sendo $\hat{\phi}$ a proporção de respostas SIM obtidas na amostra de dimensão ***n*** temos que um estimador não enviesado para P_A, será dado por:

$$\hat{P}_A = \frac{\hat{\phi} - (1-\theta)}{2\theta - 1}$$

sendo a variância deste estimador igual a

$$V(\hat{P}_A) = \frac{1}{(2\theta-1)^2} \times V(\hat{\phi}) = \frac{1}{(2\theta-1)^2} \times \frac{\phi(1-\phi)}{n} =$$

$$= \frac{P_A(1-P_A)}{n} + \frac{\theta(1-\theta)}{n(2\theta-1)^2}$$

↳► perca de precisão resultante do estratagema utilizado.

Se a questão A é sensível a sua contrária também é, e os inquiridos podem continuar a sentir que o seu anonimato não está garantido e a falsear as respostas...

[Simmons, Horvitz e outros , 1967]

Sugestão de modificações ao processo:

- 2ª questão respondida (Questão B) não fosse "melindrosa" e portanto, não estivesse relacionada com a primeira .

  Por exemplo, a questão B poderia ser: " Neste momento os minutos marcados no seu relógio estão entre os 0 e os 30 ?" SIM/NÃO

- Se a bola for preta vai responder SIM ou NÃO a esta questão.

Neste caso:

$$\phi = \theta \times P_A + (1-\theta) \times P_B \text{ e então } P_A = \frac{\phi - (1-\theta)P_B}{\theta}$$

e então,

$$\hat{P}_A = \frac{\hat{\phi} - (1-\theta)P_B}{\theta} \quad \text{com} \quad V(\hat{P}_A) = \frac{1}{\theta^2} \times V(\hat{\phi}) = \frac{\phi(1-\phi)}{n\theta^2}$$

↳► penalização na precisão.

Exemplo: (Brown e Harding, 1973)

Foram efectuados dois inquéritos a 320 oficiais da Marinha dos EUA, sobre consumo de drogas.

→ No 1º a questão foi posta directamente e o questionário era anónimo

→ No 2º utilizou-se a metodologia de respostas aleatórias.

Os resultados foram os seguintes:

Substância utilizada	1º Inquérito		2º Inquérito	
	%	EP	%	EP
Cannabis	5,0	1,2	9,0	4,1
Alucinantes	1,6	0,7	11,6	4,0
Anfetaminas	1,9	0,7	8,0	3,3,
Barbitúricos	0,6	0,7	7,9	3,9
Narcóticos	0,3	0,3	4,0	3,9

Mais tarde [Bourke, 1982] foram propostas extensões do método para tratar questões com várias respostas possíveis.

Pode-se aplicar igualmente o método ao estudo de características quantitativas discretas ou contínuas.

Característica Y → questão "melindrosa"

Característica X → questão "pacífica"

Para proteger o segredo das respostas as variáveis X e Y devem assumir valores no mesmo intervalo (as respostas x_i e y_i não devem ser distinguíveis pelo inquiridor).

Seleccionam-se independentemente duas amostras de dimensões n_1 e n_2, respectivamente.

Na 1ª amostra → n_1 na 2ª amostra → n_2

θ_1 - probabilidade de responder à questão referente a Y - θ_2

$1-\theta_1$ - probabilidade de responder à questão referente a X - $1-\theta_2$

Seja Z a resposta registada pelo inquiridor.

Na 1ª amostra: $Z_{1i} = \begin{cases} Y_{1i} \ \ com\ prob.\ \theta_1 \\ X_{1i} \ \ com\ "\ 1-\theta_1 \end{cases}$ e $\overline{Z}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} Z_{1i}$

$$E(\overline{Z}_1) = \theta_1 \mu_Y + (1-\theta_1)\mu_X$$

$$V(\overline{Z}_1) \approx \frac{1}{n_1}\left[\sigma_X^2 + \theta_1(\sigma_Y^2 - \sigma_X^2) + \theta_1(1-\theta_1)(\mu_Y - \mu_X)^2\right]$$

Na 2ª amostra: $Z_{2i} = \begin{cases} Y_{2i} \ \ com\ prob.\ \theta_2 \\ X_{2i} \ \ com\ "\ 1-\theta_2 \end{cases}$ e $\overline{Z}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Z_{2i}$

$$E(\overline{Z}_2) = \theta_2 \mu_Y + (1-\theta_2)\mu_X$$

$$V(\overline{Z}_2) \approx \frac{1}{n_2}\left[\sigma_X^2 + \theta_2(\sigma_Y^2 - \sigma_X^2) + \theta_2(1-\theta_2)(\mu_Y - \mu_X)^2\right]$$

Deduzem-se então os seguintes estimadores e suas variâncias:

$$\hat{\mu}_Y = \frac{(1-\theta_1)\overline{Z}_2 + (1-\theta_2)\overline{Z}_1}{\theta_2 - \theta_1} \qquad V(\hat{\mu}_Y) = \frac{(1-\theta_1)^2 V(\overline{Z}_2) + (1-\theta_2)^2 V(\overline{Z}_1)}{(\theta_2 - \theta_1)^2}$$

$$\hat{\mu}_X = \frac{\theta_2 \overline{Z}_1 + \theta_1 \overline{Z}_2}{\theta_2 - \theta_1} \qquad V(\hat{\mu}_X) = \frac{\theta_2^2 V(\overline{Z}_1) + \theta_1^2 V(\overline{Z}_2)}{(\theta_2 - \theta_1)^2}$$

Pode verificar-se que $\theta_2 = 0$, é o valor óptimo para minimizar $\boldsymbol{V(\hat{\mu}_Y)}$. Por conseguinte, a 2ª amostra deve ser consagrada essencialmente à estimação de $\boldsymbol{\mu_X}$.

Este método exige muito bons entrevistadores, aptos a convencer o inquirido do completo anonimato da sua resposta. Mesmo assim, não se consegue garantir a eliminação total da resposta voluntariamente errada.