Capítulo 1

Teoria da Amostragem

1.1 Introdução

A amostragem e em particular os processos de amostragem aplicam-se em variadíssimas áreas do conhecimento e constituem, muitas vezes, a única forma de obter informações sobre uma determinada realidade que importa conhecer.

A teoria da amostragem é assim um dos instrumentos que possibilita esse conhecimentos cientifico da realidade, onde outros processos ou métodos alternativos, por razões diversas, não se mostram adequados ou até mesmo possíveis.

A teoria da amostragem estuda as relações existentes entre uma população e as amostras extraídas dessa população. É útil para avaliação de grandezas desconhecidas da população, ou para determinar se as diferenças observadas entre duas amostras são devidas ao acaso ou se são verdadeiramente significativas.

Amostragem é o processo de determinação de uma amostra a ser pesquisada. A amostra é uma parte de elementos seleccionada de uma população estatística.

Enquanto que um **senso** envolve um exame a **todos** os elementos de um dado grupo, a amostragem envolve um estudo de apenas uma parte dos elementos. A amostragem consiste em seleccionar parte de uma população e observá-la com vista a estimar uma ou mais características para a totalidade da população.

"Para se saber se o bolo de chocolate está bom, basta comer uma fatia."

Alguns exemplos da utilização da amostragem são:

- Sondagens à opinião pública que servem para conhecer a opinião da população sobre variadas questões. As mais populares são as sondagens políticas.
- Inspecção de mercado utilizada com o intuito de descobrir as preferências das pessoas em relação a certos produtos. Um dos exemplos mais conhecidos da aplicação desta amostragem é a lista de audiências dos programas de televisão.
- Para estimar a prevalência de uma doença rara, a amostra pode ser constituída por algumas instituições médicas, cada uma das quais tem registo dos pacientes.

O censo apresenta dificuldades que tornam a amostragem um porco mais atraente. Entre as dificuldades que o senso apresenta, podem ser apresentadas as seguintes:

(i) A população pode ser infinita, neste caso o senso seria impossível;

(ii) A amostra pode ser actualizada mais facilmente que o censo;

(iii) O custo do senso pode torná-lo proibitivo;

(iv) Factores de tempo e custo podem apontar pela preferência entre uma amostra e um censo.

Porém há ocasiões em que o levantamento do censo pode ser vantajoso:

(i) Quando a população é pequena e o custo entre o censo e a amostra forem praticamente iguais;

(ii) Se o tamanho da amostra necessária tiver que ser muito grande em relação à população examinada;

(iii) Nas ocasiões em que se exige precisão completa;

(iv) Nas ocasiões em que já existe informação completa.

Os termos básicos em amostragem são:

- População - o grupo inteiro de objectos (unidades) dos quais se pretende obter informações. A população deve ser definida claramente e em termos daquilo que se pretende conhecer.
- Unidade - qualquer elemento individual da população.
- Amostra - uma parte ou subconjunto da população usada para obter informação acerca do todo.
- Variável - uma característica de uma unidade que será medida a partir daquela unidade da amostra.

1.1.1 As fases de um processo de amostragem

Depois de se identificar os dados que deverão ser recolhidos e o instrumento (questionário estruturado, por exemplo) a utilizar para essa recolha, o passo seguinte consiste em definir um processo de amostragem adequado ao tipo de dados e ao instrumento de análise.

No processo de recolha de dados é necessário desenvolver um processo sistemático que assegure a fiabilidade e comparabilidade desses dados. Mais especificamente, é necessário que se estabeleça à partida um plano de amostragem de acordo com a população alvo, com a definição da população a inquirir e com um processo adequado de administração do inquérito.

O plano de amostragem deverá começar por determinar qual o nível de extensão geográfica em que o processo de amostragem deverá ser conduzido (mundial, nacional, regional, urbano, rural, grupo de indivíduos, etc.).

A construção da amostra propriamente dita envolve várias etapas igualmente importantes e que são:

(i) A identificação da população alvo/população inquirida;

(ii) O método de selecção da amostra;

(iii) A dimensão da amostra.

A identificação da população alvo/população inquirida

A identificação da população de uma forma clara e objectiva é imprescindível, embora possa parecer demasiado óbvia em muitas circunstâncias. Designa-se por *população alvo* a totalidade dos elementos sobre os quais se deseja obter determinado tipo de informações.

Exemplo: Um estudo sobre as intenções de voto terá como população alvo todos aqueles que estão em idade e em condições de votar. No entanto, a população inquirida poderá incluir apenas aqueles que votaram nas últimas eleições.

Resumindo, a população alvo é constituída por todos os elementos sobre os quais se deseja obter um determinado conjunto de informações. No entanto, em muitas situações, não é operacional inquirir uma amostra retirada da população alvo, havendo necessidade de definir qual é a população a inquirir, não coincidente com a população alvo, e a partir da qual se retirará a amostra.

Os métodos de selecção da amostra

O objectivo geral na extracção de uma amostra é obter uma representação "honesta"da população que conduza a estimativas das características da população com "boa"precisão relativamente aos custos de amostragem, isto é, obter uma amostra representativa da população.

Existem dois grandes grupos de métodos para seleccionar/recolher amostras: os métodos aleatórios e métodos não aleatórios.

Os métodos de **amostragem não aleatória** são métodos ad-hoc de carácter pragmático ou intuitivo e são largamente utilizados, pois possibilitam um estudo mais rápido e com menores custos. Um claro inconveniente destes métodos é o facto de que a inclusão de um elemento da população na amostra é determinada por um critério subjectivo, normalmente uma opinião pessoal, um outro inconveniente é que existem elementos da população que não têm possibilidade de ser escolhidos.

Tipos de amostras não aleatórias:

(i) **Amostra intencional**: Composta por elementos da população seleccionados intencionalmente pelo investigador, porque este considera que esses elementos possuem características típicas ou representativas da

população;
Exemplo: escolha de localidades "representativas"em tempo de eleições legislativas.

(ii) **Amostra "snowball"**: Tipo de amostra intensional em que o investigador escolhe um grupo inicial de indivíduos e pede-lhes o nome de outros indivíduos pertencentes à mesma população. A amostra vai assim crescendo como uma bola de neve à medida que novos indivíduos são indicados ao investigador. É um tipo de amostragem bastante útil quando se pretende estudar pequenas população muito específicas (e.g. os "sem abrigo"), no entanto pode originar em resultados enviesados uma vez que as pessoas tendem a indicar o nome de pessoas intimas ou amigos (com comportamentos e pensamentos similares).

(iii) **Amostra por quotas**: As amostras são obtidas dividindo a população por categorias ou estratos e seleccionando um certo número (quota) de elementos de cada categoria de modo não aleatório.

(iv) **Amostra por conveniência**: Os elementos são escolhidos por conveniência ou por facilidade. Um exemplo diste tipo de amostragem é os casos em que os espectadores de um determinado programa são convidados a responder a um questionário. As amostras obtidas desta forma não são representativas da população e em geral são enviesadas.

Os métodos de **amostragem aleatória** são caracterizados por todos os elementos da população poderem ser seleccionados de acordo com uma probabilidade pré-definida e em que se podem avaliar objectivamente as estimativas das propriedades da população obtidas a partir da amostra.

Uma das vantagens da amostragem aleatória é a possibilidade de estimar as margens de erro dos resultados que são devidas à amostragem. Além disso, a amostragem aleatória evita o enviesamento das amostras que acontece (mesmo quando o objectivo não é esse) sempre que se usa a opinião e a experiência para escolher as amostras.

No entanto, deverão ser também referidas as dificuldades em recolher uma amostra aleatória. E a principal dificuldade consiste na obtenção de uma listagem completa da população a inquirir. Estas listagens são, na maioria dos casos, difíceis de conseguir, de custo elevado, demoradas na sua obtenção e nem sempre de fiabilidade aceitável.

O segundo tipo de dificuldades relaciona-se com as não respostas. Depois de definidos os respondentes, não poderão haver substituições, pelo que as não-respostas constituem uma importante fonte de enviesamento e terá de ser feito tudo para que a sua taxa seja minimizada. Todas as novas tentativas (por entrevista pessoal, telefone ou correio) para obter respostas bem sucedidas implicam aumento de custos e demora na obtenção dos resultados.

A amostragem aleatória é, sem dúvida, o processo mais caro, mas os custos tendem a tornar-se pouco importantes face à fiabilidade dos resultados obtidos.

Métodos de amostragem aleatória:

(i) **Amostragem aleatória simples**
Uma amostra aleatória simples de n elementos de uma população de N elementos é um subconjunto de n elementos distintos da população, extraídos de modo que qualquer das $\binom{N}{n}$ amostras possíveis tem igual probabilidade, $1/\binom{N}{n}$ de ser seleccionada.

A amostragem aleatória simples pode ser feita com reposição (caso em que cada elementos da população pode entrar mais do que uma vez na amostra) ou sem reposição (caso em que cada elemento da população só pode entrar uma vez na amostra).

Este tipo de amostra é muito dispendioso, e muitas vezes impraticável por exigir a listagem e enumeração de toda a população, daí ser poucas vezes adoptado. Mas se a população for pequena ou se existirem listas com os elementos da população, este método mostra-se bastante útil.

(ii) **Amostragem Casual sistemática**
Este método é também chamado quasi-aleatório por não dar a todas as amostras que se podem retirar de uma população a mesma probabilidade de ocorrência. Para aplicação deste método é necessário calcular o rácio $K = \frac{N}{n}$. Em seguida, escolhe-se aleatoriamente um número, no intervalo $[1, K]$, que servirá como ponto de partida e primeiro elemento da amostra. Adicionando ao primeiro valor obtido o rácio K (arredondando o resultado por defeito), obtém-se o segundo elemento

e a adição sucessiva do mesmo rácio permite encontrar os restantes elementos da amostra. Como se verifica, apenas o primeiro elemento é escolhido aleatoriamente enquanto que os restantes são determinados de modo sistemático pelo rácio.

Por exemplo, se $K = 2$, então a dimensão da amostra será constituída por metade (50%) da dimensão da população. Se $K = 20$, então a amostra será apenas 5% da população.

As empresas que executam estudos de mercado utilizam frequentemente o método denominado *Random Route*, que mais não é do que um processo de amostragem sistemática, já que partem de um ponto de partida escolhido aleatoriamente, seguindo depois um itinerário obtido com intervalos sistemáticos (inquéritos de porta a porta, por exemplo).

(iii) **Amostragem estratificada**
Este método consiste em dividir a população em grupos *relativamente* homogéneos e mutuamente exclusivos, chamados estratos, e em seleccionar amostras aleatórias simples em independentes de cada estrato. Se o número de elementos de cada amostra estiver de acordo com a proporção do estrato na população, as observações podem ser misturadas para se obter os resultados globais. Se, no entanto, todas as amostras tiverem o mesmo número de elementos, os resultados de cada estrato têm que ser pesados pela proporção desse estrato na população.

A estratificação de uma população faz sentido quando é possível identificar sub-populações que variam muito entre si no que diz respeito à variável em estudo, mas que variam pouco dentro de si. Nestas condições, uma amostra estratificada pode fornecer resultados mais precisos do que uma amostra simples extraída do conjunto da população.

Esta eficiência será ainda mais importante se a variável a ser estratificada se encontrar correlacionada com várias outras variáveis como por exemplo idade, sexo, rendimento, status, área geográfica, etc., o que permitirá estratificar simultaneamente segundo várias variáveis, desde que se assegure uma adequada representatividade dos estratos existentes na população.

(iv) **Amostragem por *clusters***

Tal como na amostragem estratificada, na amostragem por *clusters*, a população é dividida em grupos, ou *clusters*. Este tipo de amostragem torna-se particularmente útil quando a população se encontra dividida num reduzido número de grupos, caracterizados por terem uma dispersão idêntica à população total, isto é, os grupos deverão, tanto quanto possível, ser "microcosmos"da população a estudar. Primeiro, seleccionam-se aleatoriamente alguns dos grupos e em seguida, incluem-se na amostra todos os indivíduos pertencentes aos grupos seleccionados. Trata-se de um processo amostral casual simples em que cada unidade é o *cluster*.

Neste tipo de amostragem exige apenas que se disponha de uma listagem dos grupos (de indivíduos ou elementos da população) e não uma listagem completa dos elementos da população, como é o caso das amostragens anteriores.

Um exemplo deste tipo de amostragem é o caso em que se pretende fazer uma sondagem de opinião aos alunos de uma escola (população), da qual apenas se dispõe de uma listagem das turmas (grupos de alunos). Uma amostra por *clusters* obtém-se seleccionando uma amostra aleatória de turmas e inquirindo, dentro de cada turma escolhida, todos os alunos.

(v) **Amostragem multi-etapas**

O primeiro passo deste tipo de amostra é idêntico ao anterior. A população encontra-se dividida em vários grupos e seleccionam-se aleatoriamente alguns desses grupos. No passo seguinte, também os elementos de cada grupo são escolhidos aleatoriamente. Este processo pode multiplicar-se am mais de duas etapas se os grupos estiverem divididos em sub-grupos.

Um exemplo deste tipo de amostragem é o caso de uma sondagem de opinião aos alunos do ensino secundário em que se pode começar por seleccionar aleatoriamente algumas direcções escolares. Em seguida, de cada uma delas, seleccionar aleatoriamente algumas escolas, de cada uma das escolas escolhidas seleccionar aleatoriamente algumas turmas e, finalmente, de cada uma das turmas escolhidas seleccionar aleatoriamente alguns alunos. Este exemplo consiste em 4 etapas.

Como desvantagem deste método adiante-se de que os possíveis erros de amostragem se podem multiplicar, dado que ao longo deste processo se vão utilizando várias sub-amostras com a possibilidade de erros de

amostragem em cada uma delas.

(vi) **Amostragem multi-fásica**
Este processo de amostragem não deve ser confundido com o processo de amostragem multi-etapas. No primeiro processo as unidades amostrais variam de uma etapa para outra. No exemplo referido no ponto anterior, as unidades amostrais eram, sucessivamente, as direcções escolares, as escolas, as turmas e os alunos, enquanto que na amostragem multi-fásica se define sempre a mesma unidade amostral em todas as fases de extracção da amostra.

Neste caso, em cada fase da amostragem, consideram-se sempre os elementos da população, obtendo-se de alguns mais informações do que de outros. Na primeira fase, recolhem-se dados sobre determinadas características dos respondentes - por exemplo, o seu comportamento e frequência quanto ao consumo de determinado produto, variáveis demográficas, tamanho das empresas, a sua disponibilidade para responder novamente a um inquérito. Esta informação pode ser usada para a definição de uma listagem dos possíveis respondentes à segunda fase do inquérito. É então retirada desta listagem uma segunda amostra que responderá a um questionário com um nível de profundidade mais elevado.

Deste modo, nem todos os inquiridos respondem a todas as questões, isto permite reduzir os custos e permite ainda que a amostra principal seja utilizada como base de amostragem para amostragens seguintes.

1.1.2 Os conceitos principais da amostragem aleatória

O nosso interesse centra-se nos valores tomados por uma variável aleatória Y para os vários elementos de uma população e, em medidas globais dessa variável na população. Se a população tiver dimensão N, podemos representar-la por

$$Y_1, Y_2, \ldots, Y_N$$

sendo estes valores de Y designados para os diferentes membros da população.

Estamos interessados em características da população definidas relativamente a Y. As que são estudadas mais usualmente são:

(i) O total da população, $Y_T = \sum_{i=1}^{N} Y_i$;

(ii) A média da população, $Y_T = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{Y_N}{N}$;

(iii) A proporção, P, de membros da população que pertencem a determinada categoria de classificação da variável Y. Por exemplo, num estudo sobre hábitos de condução num adulto, P poderá representar a proporção de condutores que dirigem mais de 10 Km por dia.

O objectivo de um estudo por amostragem é estimar uma ou mais dessas categorias a partir da informação contida na amostra de $n(\leq N)$ membros da população. Suponha-se que os valores de Y para os membros da amostra são designados por

$$y_1, y_2, \ldots, y_N$$

onde cada y_i é um dos valores Y_j da população.

Terminologia

O quociente entre a dimensão da amostra e a dimensão da população

$$f = \frac{n}{N}$$

é chamado de **fracção amostral**.

Para estimar Y_T, $\bar{Y}$ ou P, é necessário calcular algumas medidas que sumariem a informação contida na amostra. Para estimar $\bar{Y}$ é intuitiva a utilização da média amostral

$$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{n}$$

Uma parte importante no processo de amostragem é como determinar as propriedades dos estimadores obtidos (e.g. o estimador para a média amostral dado pela equação anterior). Uma possibilidade é tentar descobrir como é que os valores de $\bar{y}$ variam relativamente a $\bar{Y}$ em diferentes situações quando se considera o procedimento amostral no mesmo problema. No entanto, para determinar as propriedades de tais estimadores, tem que se ter em conta o mecanismo aleatório de extracção de amostras.

Em termos genéricos, depois de especificar o tamanho da amostra, n, consideram-se todas as possíveis amostras de dimensão n que podem ser formadas a

partir da população, $S_1, S_2, \ldots$. Um **esquema de amostragem aleatório** é definido pela associação de uma probabilidade π_i a cada S_i, isto é, $\pi_i = P(\text{extrair a amostra } S_i)$, e escolha de uma amostra particular S de acordo com esta distribuição de probabilidade. São vastas as possibilidades para os esquemas de amostragem aleatória, correspondendo a diferentes funções de probabilidade $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots\}$ sobre o conjunto das possíveis amostras, $\{S_1, S_2, \ldots\}$.

Vamos considerar alguns dos esquemas de amostragem mais utilizados e compará-los em termos de custos e eficiência para a estimação de $\bar{Y}$, Y_T, etc.

Suponha-se que θ é uma característica da população (pode ser Y_T) e que se vai escolher uma função da amostra, $\widetilde{\theta}(S)$, para a estimar. $\widetilde{\theta}$ é designado, como usualmente, estatística ou estimador. Podem-se estudar as propriedades dos estimadores em relação à distribuição amostral de $\widetilde{\theta}$ induzida pela distribuição de probabilidade, $\boldsymbol{\pi}$. Diferentes valores de $\widetilde{\theta}$ vão ser obtidos para diferentes amostras, com probabilidades dadas por $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots\}$.

Enviesamento

Um possível critério para analisar se o esquema de amostragem é "representativo"é verificar que $\widetilde{\theta}$ é não enviesado (centrado), isto é,

$$E_\pi[\widetilde{\theta}(S)] = \theta$$

onde E representa o valor esperado.

Precisão

Usualmente o estimador $\widetilde{\theta}$ tem, pelo menos em amostras grandes, distribuição aproximadamente normal. É razoável estabelecer a precisão ou eficiência de um estimador centrado através da variância,

$$Var[\widetilde{\theta}(S)] = E_\pi\{[\widetilde{\theta}(S) - \theta]^2\}.$$

Quanto mais pequena for a variância, mais preciso é o estimador. Se, para uma dada dimensão amostral, um estimador centrado tiver menor variância do que outro, diz-se que ele é **mais eficiente**. Pode-se, assim, comparar estimadores respeitantes ao mesmo ou a diferentes esquemas de amostragem

aleatória.

O maior objectivo da teoria da amostragem é implementar esquemas de amostragem que sejam mais económicos e fáceis de implementar, e que conduzem a estimadores centrados com variância mínima.

Em geral, o factor $Var[\widetilde{\theta}(S)]$ decresce com o aumento da dimensão da amostra, mas os custos aumentam. O ideal é encontrar um ponto de equilíbrio. Têm que se comparar os esquemas de amostragem para determinar qual deles permite obter um estimador centrado com menor variância para um dado custo ou para uma dada dimensão da amostra.

1.2 Amostragem Aleatória Simples

A forma mais básica de amostragem aleatória é a amostragem aleatória simples que é relativamente simples de utilizar do ponto de vista estatístico e serve também de base a para esquemas de amostragem mais complexos como a amostragem aleatória estratificada e a amostragem aleatória por grupos. As propriedades dos estimadores obtidos a partir de amostras aleatórias simples são facilmente demonstrados.

1.2.1 O procedimento de Amostragem Aleatória Simples

Se a população tiver dimensão N, e quisermos uma amostra aleatória simples de dimensão n, esta amostra é escolhida aleatoriamente das $\binom{N}{n}$ amostras distintas possíveis, em cada uma das quais nenhum dos elementos da população é incluído mais de uma vez. Isto é o mesmo que dizer que cada uma das $\binom{N}{n}$ amostras possíveis tem a mesma probabilidade $\binom{N}{n}^{-1}$ de ser escolhida.

Para produzir uma amostra aleatória simples de dimensão n (amostra aleatória sem reposição de n elementos da população) deve-se proceder do seguinte modo. Suponha-se que este método de extracção sequencial sem reposição produz n elementos (distintos) da população cujos valores são

$$y_1, y_2, \ldots, y_n$$

onde y_i se refere ao i-ésimo elemento, $i = 1, \ldots, n$.

A probabilidade de obter esta sucessão ordenada é

$$\frac{1}{N}.\frac{1}{N-1}\cdots\frac{1}{N-n+1} = \frac{(N-n)!}{N!}$$

Mas, qualquer ordenação de $y_1, y_2, \ldots, y_n$ corresponde à mesma escolha de n elementos distintos da população (isto é, corresponde à mesma amostra). Existem $n!$ ordenações possíveis. Assim, a probabilidade de obter uma amostra particular de n elementos (independente da ordem) é dada por

$$\frac{n!(N-n)!}{N!} = \binom{N}{n}^{-1}.$$

Existem $\binom{N}{n}$ amostras distintas e são igualmente prováveis, isto é, são amostras aleatórias simples.

A escolha de uma observação individual na amostra é conseguido em cada etapa por um mecanismo aleatório aplicado aos restantes membros da população, por exemplo, utilizando uma tabela de números aleatórios.

Exemplo 1.2.1: Quer-se extrair uma amostra aleatória simples de 5 elementos de 25. Primeiro deve-se numerar a população de 0 a 24, depois procurar numa tabela de números aleatórios os primeiros pares de números menores que 25, obtendo assim os 5 elementos da população que devem ser seleccionados. Não esquecer de medir o respectivo valor desses elementos na variável em estudo, nem de ignorar os que foram seleccionados anteriormente na procura na tabela de números aleatórios. Para amostras e populações grandes, esta tarefa de escolher a amostra a partir de uma tabela de números aleatórios pode ser demasiado morosa.

Variância

A variância de uma população finita $Y_1, Y_2, \ldots, Y_N$ é dada por

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2.$$

Na amostragem aleatória simples pode-se definir o valor esperado de y_i, a i-ésima observação na amostra, isto é,

$$E[y_i] = \sum_{j=1}^{N} Y_i P(y_i = Y_j) = \frac{1}{N} \sum_{j=1}^{N} Y_j = \bar{Y}.$$

O resultado que diz que $P(y_i = Y_j) = \frac{1}{N}$ é devido ao facto de que o número de amostras em que $y_i = Y_j$ ser de $\frac{(N-1)!}{(N-n)!}$, e cada uma tem probabilidade de $\frac{(N-n)!}{N!}$.

Facilmente se verifica que

$$E[y_i^2] = \frac{1}{N} \sum_{j=1}^{N} Y_j^2,$$

e

$$E[y_i y_j] = \frac{2}{N(N-1)} \sum_{r<s} Y_r Y_s \qquad (i \neq j)$$

Assim, a variância e covariância de y_i são dadas por

$$\begin{aligned} Var[y_i] &= E[(y_i - \bar{Y})^2] \\ &= E[y_i^2] - \bar{Y}^2 \\ &= \frac{(N-1)\sigma^2}{N} \end{aligned}$$

e

$$\begin{aligned} Cov[y_i, y_j] &= E\{(y_i - \bar{Y})(y_i - \bar{Y})\} \\ &= E[y_i y_j] - \bar{Y}^2 \\ &= \frac{1}{N(N-1)} \left[\left(\sum_{j=1}^{N} Y_j \right)^2 - \sum_{j=1}^{N} Y_j^2 - N(N-1)\bar{Y}^2 \right] \\ &= -\frac{\sigma^2}{N}. \end{aligned}$$

Pode-se assim concluir que existe uma pequena e negativa correlação entre as potenciais observações amostrais.

Pode-se, agora, proceder ao estudo do estimador da média da população.

1.2.2 Estimação da média, $\bar{Y}$

Um estimador de $\bar{Y}$, baseado numa amostra aleatória simples de dimensão n, imediatamente intuitivo é a média amostral,

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Facilmente se verifica que $\bar{y}$ é um estimador centrado de $\bar{Y}$, pois

$$E[\bar{y}] = \frac{1}{n} E\left[\sum_{i=1}^{n} y_i \right] = \frac{n\bar{Y}}{n} = \bar{Y}.$$

Além disso,

$$Var[\bar{y}] = \frac{(1-f)\sigma^2}{n}, \tag{1.1}$$

em que $f = \frac{n}{N}$ é a fracção amostral.

A variância amostral de $\bar{y}$ é reduzida por um factor $f = \frac{n}{N}$, fracção de amostragem, comparado com o resultado análogo para uma população infinita. Este efeito é conhecido como correcção de população finita (c.p.f.). Se o valor da fracção amostral for muito pequeno, a c.p.f. tem pouca importância e pode ser ignorada. Empiricamente, pode-se ignorar a c.p.f. se f é menor ou igual a 0.05. A consequência deste procedimento é obter-se uma variância um pouco maior para o estimador $\bar{y}$.

Terminologia

O erro padrão (*standard error*) de $\bar{y}$ é dado por $[Var(\bar{y})]^{1/2}$.

Pode-se dizer que $\bar{y}$ é um estimador centrado de $\bar{Y}$ e (1.1) permite-nos comparar a eficiência de diferentes estimadores de $\bar{Y}$ baseados em amostragem aleatória simples ou amostras obtidas por outros processos de amostragem.

Além disso, $\bar{y}$ é um estimador consistente de $\bar{Y}$ no caso de populações finitas, isto é, quando $n \to N$, $\bar{y} \to \bar{Y}$.

Quanto à questão de saber como é que $\bar{y}$ se compara com outros possíveis estimadores de $\bar{Y}$, num esquema de amostragem aleatória simples, pode ser apresentada a seguinte propriedade, facilmente demonstrável:

Propriedade: A média amostral, $\bar{y}$, é o melhor (com menor variância) estimador linear centrado de $\bar{Y}$ baseado numa amostra aleatória de dimensão n.

1.2.3 Amostragem Aleatória com reposição

Observe-se como os resultados diferem se for utilizado um método de amostragem aleatório simples, mas agora com reposição, para obtenção de uma amostra aleatória de dimensão n de uma população de dimensão N.

A amostragem aleatória simples com reposição de uma população finita é um método de mostragem em que cada elemento Y_i da amostra $Y_1, Y_2, \ldots, Y_n$ é escolhido aleatoriamente entre todos os N elementos da população $y_1, y_2, \ldots, y_n$, e de forma que todos os elementos da população tenham a mesma probabilidade de serem escolhidos, isto é, $P(Y_i = y_k) = \frac{1}{N}, i = 1, 2, \ldots, n; k = 1, 2, \ldots, N$. Isto corresponde a extrair uma amostra aleatória de dimensão n de uma uma distribuição uniforme discreta no conjunto dos pontos $Y_1, Y_2, \ldots, Y_N$.

Observe-se que, neste caso, cada elementos da amostra é estatisticamente independente dos restantes, e todos os elementos são identicamente distribuídos e têm a mesma distribuição de probabilidade da população.

Verifica-se facilmente que:

- $E(y_i) = \bar{Y},\ i = 1, 2, \ldots, n;$
- $E(y_i^2) = \frac{1}{N} \sum_{i=1}^{N} Y_j^2;$
- $Var(y_i) = \frac{N-1}{N} \sigma^2.$

Se se considerar a média amostral $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ como sendo o estimador de $\bar{Y}$, tem-se que

- $E(\bar{y}) = \bar{Y};$
- $Var(\bar{y}) = \frac{1}{n} \left(1 - \frac{1}{N}\right) \sigma^2.$

Compare-se este último resultado para a variância com a expressão (1.1), $\frac{1}{n} \left(1 - \frac{n}{N}\right) \sigma^2$, para o caso da amostragem aleatória simples (sem reposição).

O estimador $\bar{y}$ de $\bar{Y}$ referente à amostragem aleatória com reposição é menos eficiente que o mesmo estimador referente à amostragem aleatória simples, uma vez que $1 - f < 1 - \frac{1}{N}$ para $n > 1$. A sua eficiência relativa é dada por $\frac{N-n}{N-1}$.

1.2.4 Estimação da variância σ^2

A expressão (1.1) para $Var(\bar{y})$ é utilizada de três formas:

(i) para estabelecer a precisão do estimador $\bar{y}$ de $\bar{Y}$;

(ii) para comparar $\bar{y}$ com outros estimadores de $\bar{Y}$;

(iii) Para determinar a dimensão da amostra necessária para obter a precisão do estimador $\bar{y}$ pretendida.

Normalmente, não se conhece o verdadeiro valor de σ^2, como tal é necessário estimá-lo a partir da amostra. Considerando a amostra aleatória simples $y_1, y_2, \ldots, y_n$, utiliza-se, como habitualmente,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Note-se que s^2 é um estimador centrado de σ^2, isto é, $E(s^2) = \sigma^2$.

Relativamente aos pontos (i) e (ii), pode-se substituir a variância desconhecida da população, σ^2, em (1.1) pelo seu estimador centrado s^2, obtendo-se assim um estimador centrado de $Var(\bar{y})$ dado por

$$s^2(\bar{y}) = (1-f)\frac{s^2}{n}.$$

Em algumas situações, a estimação de σ^2 é útil, por si só, e tal estimação pode ser feita utilizando os estimador s^2. Mas quanto ao problema referido em (iii), em se quer determinar a dimensão da amostra necessária para obter a precisão pretendida, o estimador s^2 não tem relevância porque ainda não se dispõe da amostra para o calcular. Como tal, tem que se determinar a dimensão da amostra requerida antes de efectuar o processo de amostragem. Posteriormente ver-se-á como realizar este processo.

1.2.5 Intervalo de confiança para $\bar{Y}$

Para se obter um intervalo de confiança para $\bar{Y}$ é necessário que se conheça a sua distribuição. Como se está perante um caso de amostragem, o que se pretende é a distribuição por amostragem, e, a forma de a obter é utilizar um caso análogo ao Teorema do Limite Central para populações finitas que permite concluir que a média amostral, $\bar{y}$, de uma amostra aleatória simples tem aproximadamente distribuição normal,

$$\bar{y} \sim N\left(\bar{Y}, (1-f)\frac{\sigma^2}{n}\right) \tag{1.2}$$

Esta suposição é usualmente bastante razoável, mesmo se existe simetria na população. Uma regra empírica para a utilização desta aproximação de $\bar{y}$ é que a dimensão da amostra, n, satisfaça

$$n > 25G_1^2$$

onde

$$G_1 = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(Y_i - \bar{Y})^3$$

Note-se que para populações finitas G_1 é o análogo ao coeficiente de assimetria de Fisher. Além disso, a função de amostragem, $f = \frac{n}{N}$ não deve ser muito grande.
Quando esta aproximação é apropriada, pode-se utilizar a distribuição normal para realizar inferências sobre $\bar{Y}$. Um intervalo de confiança a $100(1-\alpha)\%$ para $\bar{Y}$ pode ser escrito da seguinte forma

$$\left]\bar{y} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sigma\sqrt{\frac{1-f}{n}}; \bar{y} + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sigma\sqrt{\frac{1-f}{n}}\right[; \qquad (1.3)$$

Mas na prática, o valor de σ^2 não é conhecido e tem que se utilizar a sua estimativa, s^2. Isto é razoável se o valor de n for suficientemente grande.

No caso do valor de n não ser grande (se $n \leq 40$) pode-se utilizar a distribuição t de Student e o intervalo de confiança a $100(1-\alpha)\%$ para $\bar{Y}$ é dado por

$$\left]\bar{y} - t_{n-1,1-\alpha/2}.s.\sqrt{\frac{1-f}{n}}; \bar{y} + t_{n-1,1-\alpha/2}.s.\sqrt{\frac{1-f}{n}}\right[\qquad (1.4)$$

onde $t_{n-1,1-\alpha/2}$ é o quantil de probabilidade $1-\alpha/2$ da distribuição t de Student com $n-1$ graus de liberdade.

Geralmente, os inquéritos por amostragem são relativos a populações muito grandes ($N = 10000$ ou mais) com dimensões amostrais substanciais ($n = 100$ ou mais). Assim, usualmente utiliza-se a forma do intervalo de confiança (1.3)

substituindo σ^2 por s^2.

Exemplo: Para investigar a taxa de absentismo não relacionado com feriados ou férias, mum sector da indústria foi realizado um inquérito. Foi recolhida uma amostra aleatória de 1000 indivíduos de um total de 36000 trabalhadores, aos quais foi questionado quantos dias tinham faltado ao trabalho nos 6 meses anteriores. Os resultados obtidos foram os seguintes:

Número de faltas	0	1	2	3	4	5	6	7	8	9
Número de trabalhadores	451	162	187	112	49	21	5	11	2	0

Para estimar o número médio, $\bar{Y}$ de faltas, dadas pelos empregados deste sector, nos últimos 6 meses pode-se utilizar a média amostral

$$\bar{y} = 1.296$$

A variância amostral é dada por

$$s^2 = 2.397$$

Utilizando uma aproximação à distribuição normal para a média, $\bar{y}$, obtém-se um intervalo de confiança a 95% para $\bar{Y}$ dado por

$$\Big]1.296 \pm; 1.96\sqrt{2.397}\sqrt{(1 - 1000/36000)/1000}\Big[=]1.201; 1.391[$$

(ou $]1.200; 1.392[$ se se ignorar a c.p.f. uma vez que $f = \frac{n}{N} = \frac{1}{36} = 0.028 < 5\%$)

Note-se que a distribuição dos valores de Y na população é altamente assimétrica. Este facto afecta a qualidade da aproximação normal, mas a dimensão elevada da amostra e da população compensa esse facto.

1.2.6 Escolha da dimensão da amostra

É evidente que um aumento da dimensão da amostra conduzirá a um aumento da precisão de $\bar{y}$ como estimador de $\bar{Y}$.Contudo os custos de amostragem também irão aumentar e existem limites para aquilo que podemos gastar. Uma amostra demasiado grande implica um desperdício de esforço; uma amostra demasiado pequena produzirá uma estimação de precisão inadequada. O ideal será estabelecermos a precisão desejada, ou o gasto máximo que podemos realizar, e escolher a dimensão da amostra de acordo com estas

restrições.

Para alcançar este objectivo é necessário ter em conta um vasto leque de considerações:

- Conhecer o custo de amostragem para dada situação;
- Saber como aferir da precisão dos estimadores;
- Saber como equilibrar as necessidades em relação a várias características da população que estejam a ser estimadas (características de interesse).
- Como lidar com o desconhecimento de alguns parâmetros da população (e.g. a variância da população) que podem afectar a precisão dos estimadores.

Vai-se considerar apenas um caso simples. Vai-se assumir que o objectivo é estimar apenas uma característica, a média da população, $\bar{Y}$, utilizando a média $\bar{y}$ obtida a partir de uma amostragem aleatória simples, e impondo que a probabilidade da diferença absoluta entre $\bar{y}$ e $\bar{Y}$ ser superior a um dado valor não exceda um certo nível. Não fazemos quaisquer considerações sobre custos embora, se os custos de amostragem forem proporcionais à dimensão da amostra, o objectivo de redução ao mínimo custo seja alcançado do mesmo modo.

Suponhamos que procuramos encontrar o valor mínimo de n que assegura que

$$P\left(|\bar{Y} - \bar{y}| > d\right) \leq \alpha \tag{1.5}$$

para valores especificados de d (tolerância) e (pequeno) α (risco de não respeitar essa tolerância). (1.5) pode ser escrito como

$$P\left(\frac{|\bar{Y} - \bar{y}|}{\sigma\sqrt{(1-f)/n}} > \frac{d}{\sigma\sqrt{(1-f)/n}}\right) \leq \alpha, \tag{1.6}$$

assim, utilizando a aproximação à distribuição normal de $\bar{y}$ pode-se escrever

$$\frac{d}{\sigma\sqrt{(1-f)/n}} \geq \Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big) \tag{1.7}$$

ou ainda

$$n \geq N \left[1 + N \left(\frac{d}{\sigma . \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)} \right)^2 \right]^{-1} \tag{1.8}$$

A inequação (1.6), declara de modo equivalente que

$$Var(\bar{y}) \leq \left(\frac{d}{\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)} \right)^2 = V, \tag{1.9}$$

e portanto a desigualdade (1.8) pode ser escrita como

$$n \geq \frac{\sigma^2}{V} \left[1 + \frac{1}{N} \frac{\sigma^2}{V} \right]^{-1}, \tag{1.10}$$

Verificamos assim que, como primeira aproximação para a pretendida dimensão da amostral, podemos considerar

$$n_0 = \frac{\sigma^2}{V}. \tag{1.11}$$

Contudo esta expressão avalia por excesso a dimensão da amostra, especialmente se a fracção de amostragem $f = \frac{n_0}{N}$ for substancial. Se tal acontecer, é necessário diminuir a nossa aproximação e, em vez de n_0, considerar

$$n = n_0 \left(1 + \frac{n_0}{N} \right)^{-1} \tag{1.12}$$

Tudo isto pressupõe naturalmente que σ^2 é conhecido. Na prática isso não acontece, como tal é necessário estimar a dimensão da amostra requerida, n quando σ^2 é desconhecido. Existem basicamente 4 formas de o fazer:

(i) **A partir de estudos piloto**: Muitas vezes é possível fazer um estudo piloto antes do inquérito principal. Se tal for feito os resultados dão alguma indicação sobre o valor de σ^2 a utilizar na escolha da dimensão da amostra. No entanto, esta estimativa poderá ser bastante enviesada uma vez que os estudos piloto incidem, em geral, sobre uma parte da população apenas.

(ii) **A partir de inquéritos anteriores**: É bastante comum repetir estudos anteriores para estudar características similares em populações similares, especialmente em áreas como a educação, a medicina ou sociologia. A medida para a variância, σ^2 nesses estudos anteriores poderá ser utilizada no novo estudo de modo a determinar a dimensão da amostra, no entanto é necessário cautela ao extrapolar de uma população para a outra.

(iii) **A partir de uma amostra preliminar**: Esta é a abordagem mais objectiva e mais indicada, mas pode não ser admissível em termos administrativos ou de custos. O procedimento consiste em recolher uma amostra aleatória simples de pequena dimensão, n_1, e utilizar a variância amostral, s_1^2 para estimar a variância, σ^2. Com esta estimativa de σ^2 calculamos o valor mínimo para n, após o qual se recolhem mais $(n - n_1)$ observações dos restantes elementos da população.
Com este procedimento, e se for razoável ignorar a correcção de população finita (c.p.f.), a dimensão da amostra, n, deverá ser igual a

$$\left(1 + \frac{2}{n_1}\right) \frac{s_1^2}{V}$$

Este processo de amostragem é um caso de amostragem em 2 fases.

(iv) **A partir de considerações práticas acerca da estrutura da população**: Ocasionalmente temos algum conhecimento sobre a estrutura da população de que pode dar indicação sobre o valor de σ^2. Por exemplo, considerem-se o número de "gralhas"em livros de uma dada editora (aproximadamente do mesmo tamanho ou mum número prefixado de páginas) num certo período de tempo, ou o número de falhas que ocorrem numa marca de cassetes de vídeo no primeiro ano de uso. Em ambos os casos se pode admitir que os valores da variável em estudo, Y, seguem uma distribuição de Poison, sendo então plausível considerar que σ^2 e $\bar{Y}$ sejam aproximadamente iguais. Logo, qualquer informação sobre $\bar{Y}$ pode ser utilizada para estimar σ^2 e intervir na escolha da dimensão da amostra, n.

1.2.7 Estimação do total da população, Y_T

Existem muitas situações em que é interessante estimar o total da população

$$Y_T = N\bar{Y}. \tag{1.13}$$

em vez da média da população, $\bar{Y}$. Através desta relação entre Y_T e $\bar{Y}$ podemos, facilmente deduzir as propriedades sobre estimação do total populacional.

O estimador por amostragem aleatória simples que é mais utilizado é dado por

$$y_T = N\bar{y}$$

Dos resultados anteriores, tem-se que y_T é um estimador centrado de Y_T e

$$Var(y_T) = N^2(1-f)\frac{\sigma^2}{n}.$$

y_T é o estimador linear centrado de variância mínima de Y_T baseado numa amostra aleatória simples de dimensão n.
Com as mesmas restrições relativamente à dimensão da amostra, n, e ao valor da fracção de amostragem, f, pode-se usar a aproximação à distribuição normal dada por

$$y_T \sim N\left(Y_T, \frac{(1-f)N^2\sigma^2}{n}\right)$$

para construir intervalos de confiança para Y_T. Se $n > 40$, um intervalo de confiança a $100(1-\alpha)$ para Y_T é dado por

$$\left]y_T - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sigma N\sqrt{\frac{1-f}{n}}; y_T + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sigma N\sqrt{\frac{1-f}{n}}\right[;$$

Se $n \leq 40$, é preferível utilizar o quantil $t_{n-1,1-\frac{\alpha}{2}}$ em vez do quantil $\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ da distribuição normal reduzida.

Quanto à questão da escolha da dimensão da amostra, n, tem-se em conta que

$$P\left(|y_T - Y_T| > d\right) \leq \alpha.$$

Utilizando a aproximação pela distribuição normal, vem que

$$n \geq N \left[1 + \frac{1}{N} \left(\frac{d}{\sigma . \Phi^{-1} \left(1 - \frac{\alpha}{2} \right)} \right)^2 \right]^{-1} . \tag{1.14}$$

Equivalentemente,

$$Var(y_T) \leq \left(\frac{d}{\Phi^{-1} \left(1 - \frac{\alpha}{2} \right)} \right)^2 = V$$

Assim, (1.14) pode ser escrito como

$$n \geq \frac{N^2 \sigma^2}{V} \left(1 + \frac{1}{N} \frac{N^2 \sigma^2}{V} \right)^{-1}$$

Assim, se $\frac{n\sigma^2}{V}$ é muito menor que 1, é razoável tomar

$$n_0 = \frac{n^2 \sigma^2}{V}$$

como dimensão aproximada da amostra, caso contrário deve-se utilizar

$$n_0 \left(1 + \frac{n_0}{N} \right)^{-1} .$$

1.2.8 Estimação de uma proporção, P

O objecto de um estudo de amostragem pode incidir sobre um atributo ou qualidade dos elementos de uma população, nomeadamente sobre o estudo da proporção de indivíduos da população que tem o atributo. Por exemplo a proporção de casas alugados na área da grande Lisboa. Já vimos que podemos atribuir o valor 1 aos elementos da população que têm o atributo e o valor 0 aos elementos que não têm o atributo. Do mesmo modo, a amostra vai ser constituída por 0s e 1s, isto é, $x_i = 1$ se o i-ésimo elemento da amostra tem o atributo e $x_i = 0$ se o i-ésimo elemento da amostra não tem o atributo.

Sendo assim, se r elementos da amostra tiverem o atributo, então

$$\sum_{i=1}^{n} x_i = r.$$

pelo que

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{r}{n}.$$

é a proporção de elementos da amostra que têm o atributo e que vamos representar como p.

$$p = \frac{r}{n}.$$

é o estimador de $P = \frac{R}{N}$.

Constata-se assim que o estudo da estimação de uma proporção, P, é equivalente ao estudo da estimação de um valor médio, $\bar{X}$.

Ao discutir a eficácia de p como estimador de P, estamos a discutir o uso da média de uma amostra aleatória simples como estimador da média da população. No entanto, existe neste caso a particularidade de os valores da variável X poderem ser apenas 0 e 1. Isto implica a existência de uma relação entre $\bar{X}$ (ou seja, P) e $\sigma^2_{\bar{X}}$. De facto,

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - P)^2 = \frac{NP(1-P)}{N-1} \tag{1.15}$$

uma vez que

$$\begin{aligned}
\sigma^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}(X_i - P)^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}X_i^2 - \frac{N}{N-1}P^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}X_i - \frac{N}{N-1}P^2 \\
&= \frac{NP(1-P)}{N-1}
\end{aligned}$$

Fazendo as devidas adaptações é fácil obter as propriedades do estimador p.

$$E(p) = P, \text{ isto é, o estimador é centrado.}$$

e,

$$Var(p) = (1-f)\frac{\sigma_X^2}{n} = \frac{N-n}{N-1}\frac{P(1-P)}{n}$$

Mas, como σ_X^2 é desconhecido, pode-se estimar pelo seu estimador centrado

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \cdots = \frac{n}{n-1}p(1-p),$$

e, consequentemente,

$$s^2(p) = (1-f)\frac{p(1-p)}{n-1}$$

é um estimador centrado de $Var(p)$.

1.2.9 Intervalos de confiança para P

Havendo R elementos da população com o atributo, então a probabilidade de na amostra se observarem r elementos com o atributo é

$$P(r) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}, \qquad \max(0, n-N+R) \leq r \leq \min(R, n)$$

ou seja, o número de elementos da amostra, de dimensão n, com o atributo tem distribuição de parâmetros (N, R, n). Conhecendo o modo de determinar as probabilidades podemos sempre construir intervalos de confiança para P. Contudo, se utilizarmos esta distribuição exacta, hipergeométrica, os cálculos para a obtenção dos intervalos de confiança são muito pesados.

Podemos tentar então uma primeira aproximação da distribuição hipergeométrica à distribuição binomial, que sabemos ser razoável, desde que $f = \frac{n}{N} \leq 10\%$. Assim, considerando que o número de elementos da amostra com o atributo tem distribuição aproximadamente binomial de parâmetros (n, P), é possível obter intervalos de confiança para P. Mas também neste caso os cálculos são pesados.

Resta-nos a aproximação à distribuição normal, que sabemos ser razoável se:

(i) n não muito grande relativamente a R ou a $N-R$;

(ii) $\min(np, n(1-p)) > 30$.

Verificando-se então que com

$$Var(p) \approx \frac{N-n}{N}\frac{P(1-P)}{n} = (1-f)\frac{P(1-P)}{n},$$

tem-se que

$$\frac{p-P}{\sqrt{(1-f)\frac{P(1-P)}{n}}} \tag{1.16}$$

tem distribuição aproximadamente normal reduzida.

E, sabendo que

$$P\left(\left|\frac{p-P}{\sqrt{(1-f)\frac{P(1-P)}{n}}}\right| \leq \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) = 1-\alpha, \tag{1.17}$$

o intervalo de confiança a $100(1-\alpha)\%$ para P é dado pela região entre as duas raízes da equação quadrática (em P) dada por

$$P^2\left(1+\frac{1-f}{n}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) - P\left(2p+\frac{1-f}{n}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) + p^2 = 0$$

Se n for suficientemente grande, podemos simplificar ainda mais. Substituindo $Var(p)$ pelo seu estimador centrado, $s^2(p)$, na distribuição aproximada normal de p, obtém-se o intervalo de confiança a $100(1-\alpha)\%$ para P dado por:

$$p \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{(1-f)\frac{p(1-p)}{n-1}} \tag{1.18}$$

1.2.10 Escolha da dimensão da amostra na escolha de uma proporção

Recorde-se que $Var(p) = \frac{N-n}{N-1}\frac{P(1-P)}{n} = (1-f)\frac{P(1-P)}{n}$. Claramente, esta variância é máxima para $P = \frac{1}{2}$, o que significa que, para uma dada dimensão, n, da amostra, a estimação de P é menos precisa quando P for próximo de $\frac{1}{2}$. Para $\frac{1}{4} < P < \frac{3}{4}$, $\sqrt{P(1-P)}$ (que reflecte o desvio padrão de p) apenas varia no intervalo $(0.433, 0.500)$, e a variação na precisão do estimador de p é muito pequena. É necessário que $P = 0.07$ ou $P = 0.93$ para que o desvio padrão seja reduzido para 50% do seu máximo valor.

A escolha da dimensão da amostra que assegura certos limites para o erro padrão da estimativa de P, vai ser uma vez mais equivalente à escolha da dimensão da amostra que assegura, com um probabilidade α predefinida, uma precisão, absoluta, d, ou proporcional, ξP, para o estimador p.

CASO A

Suponhamos que, para os valores pré-estabelecidos d e α (pequeno), pretendemos encontrar a dimensão da amostra que assegura que

$$P(|p - P| > d) \leq \alpha.$$

Considerando a aproximação à distribuição normal (1.16), isto é equivalente a exigirmos que

$$Var(p) = \frac{N-n}{N-1}\frac{P(1-P)}{n} \leq \left(\frac{d}{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}\right)^2$$

do que resulta

$$n \geq N\left(1 + \frac{N-1}{P(1-P)}V\right)^{-1} \quad \text{sendo} \quad V = \left(\frac{d}{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}\right)^2 \tag{1.19}$$

ou ainda

$$n \geq \frac{P(1-P)}{V}\left(1 + \frac{1}{N}\left(\frac{P(1-P)}{V} - 1\right)\right)^{-1} \tag{1.20}$$

Como primeira aproximação podemos considerar

$$n_0 = \frac{P(1-P)}{V},$$

que é a expressão obtida se ignorarmos a correcção de população finita. Se $\frac{n_0}{N}$ não for pequeno, deve-se usar a expressão mais exacta (1.19), isto é,

$$n \geq n_0 \left(1 + \frac{n_0 - 1}{N}\right)^{-1}$$

CASO B

Por vezes pretende-se uma precisão para a estimativa de P expressa em termos de proporcionalidade em relação a P, ou seja, pensar numa precisão $d = \xi P$. Isto significa desejar que o estimador tenha uma certa precisão relativa, isto é, que o erro relativo do estimador não exceda ξ com probabilidade $1 - \alpha$.

Sendo assim, para ξ e α (pequeno) pré-estabelecidos, queremos saber qual a dimensão da amostra que assegura que

$$\begin{aligned} & P(|p - P| > \xi P) \leq \alpha \\ \iff \quad & P\left(\frac{|p - P|}{\sqrt{Var(p)}} > \frac{\xi P}{\sqrt{Var(p)}}\right) \leq \alpha \end{aligned} \tag{1.21}$$

Utilizando a aproximação à distribuição normal, pode-se escrever, como anteriormente,

$$n \geq N \left(1 - \frac{N-1}{U}\right)^{-1} = U \left(1 + \frac{1}{N}(U-1)\right)^{-1} \tag{1.22}$$

sendo

$$U = \frac{1-P}{P} \left(\frac{\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)}{\xi}\right)^2$$

Pode considerar-se como primeira aproximação de n o valor

$$n_0 = U = \frac{1-P}{P} \left(\frac{\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)}{\xi}\right)^2.$$

Contudo, se $\frac{n_0}{N}$ não for pequeno é mais conveniente considerar

$$n \geq n_0 \left(1 + \frac{n_0 - 1}{N}\right)^{-1}$$

1.3 Estimadores de uma razão e de regressão

No capítulo anterior apenas se considerou a estimação de uma única característica da população com base num esquema de amostragem aleatória simples. Considerando o mesmo processo de amostragem vai-se alargar um pouco o estudo, considerando mais do que uma característica de interesse. Frequentemente, o objectivo de um inquérito por amostragem, é obter informação sobre várias características populacionais. Assim, está-se muitas vezes perante dados multivariados que dizem respeito a várias medidas da população, representadas pelas variáveis X, Y, Z, ...

A estimação simultânea de várias características populacionais explorando a estrutura de correlação da população multivariada não é a parte principal deste estudo. Contudo, vai ser abordada com algum detalhe uma extensão da situação univariada. Trata-se do caso bivariado, em que se observam simultaneamente duas variáveis, X e Y. Vão ser discutidas duas situações, com objectivos distintos, mas que envolvem considerações estatísticas semelhantes:

(i) como estimar a razão de duas características populacionais, por exemplo $\frac{Y_T}{X_T}$,

(ii) como estimar eficientemente uma característica populacional relativamente a uma variável de estudo, Y, por exemplo $\bar{Y}$ ou Y_T, explorando a associação existente entre as variáveis X e Y observadas anteriormente.

1.3.1 Estimação de uma razão

Em várias situações pretende-se estimar uma razão de duas características populacionais: os totais ou as médias de duas variáveis em estudo X e Y. Estamos interessados em estimar a quantidade

$$R = \frac{Y_T}{X_T} = \frac{\bar{Y}}{\bar{X}}$$

que será designada por razão populacional.

O interesse em estimar R pode surgir de duas formas. Esta razão pode ter interesse em si mesmo, por exemplo, pode-se querer estimar a proporção de terra arável cultivada de centeio numa determinada região geográfica. Para isso, recolhe-se uma amostra das quintas da região e regista-se para cada

uma delas a área total e a área utilizada no cultivo do centeio. Se se designar essas áreas por X_i e Y_i para as diferentes quintas da região, o que se quer estimar é $R = \frac{T_T}{X_T}$.

Alternativamente, o interesse por uma razão, R, pode surgir devido a conveniências administrativas na montagem de um esquema de amostragem que seja viável. Suponha-se que se queria estimar o rendimento anual médio por pessoa, ou número médio de carros por pessoa, para a população adulta residente numa determinada região geográfica. Poder-se-ia pensar em recolher uma amostra aleatória simples de indivíduos adultos, registar o seu rendimento anual ou o número de carros que possui (predominantemente 0 e 1) e utilizar a média amostral, em cada caso, para estimar a correspondente média populacional. Mas, pode não ser fácil obter uma amostra aleatória dos adultos, por exemplo devido à dificuldade em ter acesso à população ou outras quantidades de interesse. Pode ser mais simples utilizar unidades de amostragem maiores, como por exemplo os agregados familiares. Neste caso, passa a ter interesse estimar razões, em vez de médias. O rendimento anual médio pode ser agora interpretado como a razão entre o rendimento total anual dos agregados familiares, Y_T e o número total de adultos da população, X_T, sendo ambas as características estimadas a partir de uma amostra de agregados familiares. O raciocínio é análogo para o número médio de carros por pessoa.
Note-se que, nestes exemplos, se utilizam grupos de indivíduos como unidades de amostragem para estudar características por indivíduo.

Assim, estamos interessados em estimar a razão $R = \frac{Y_T}{X_T}$, com base numa amostra aleatória simples $(y_1, x_1), \ldots, (y_n, x_n)$ dos valores de uma população bivariada $(Y_i, X_i), i = 1 \ldots, N$.

Existem várias abordagens possíveis para a estimação de R. Duas abordagens imediatas são, utilizar a razão média da amostra ou a razão das médias amostrais. Mais especificamente, elas são

$$r_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i}$$

e

$$r_2 = \frac{\bar{y}}{\bar{x}} = \frac{y_T}{x_T}$$

respectivamente.

Muitas vezes, os valores das variáveis X e Y estão correlacionados. Por exemplo, se Y representar o gasto de um agregado familiar em alimentação e X representar o rendimento do agregado familiar, é natural esperar uma correlação positiva entre as variáveis X e Y. è também claro que a presença ou ausência de correlação entre as duas variáveis vai afectar as propriedades dos estimadores r_1 e r_2. Por exemplo, se existir uma correlação positiva elevada entre X e Y, as razões individuais $\frac{Y_i}{X_i}$ vão variar pouco, comparado com uma situação em que as variáveis não estão correlacionadas (supondo variâncias, σ_Y^2 e σ_X^2, iguais para ambas as situações) e este facto vai-se reflectir na precisão dos estimadores.

Estimador r_1

Apesar do seu caracter intuitivo, r_1 não é muito utilizado como estimador da razão populacional R. r_1 é um estimador enviesado e, quer o viés quer o erro quadrático médio, podem ser elevados relativamente aos valores de outros estimadores, em particular de r_2. O viés deste estimador pode ser calculado rapidamente.

Considere-se a população de valores $R_i = \frac{Y_i}{X_i}$. A média populacional é dada por

$$\bar{R} = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i}{X_i}$$

e a variância é

$$\sigma_R^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(R_i - \bar{R}\right)^2$$

Desde que r_1 seja a média de uma amostra aleatória simples (isto é, $r_1 = \bar{r}$), r_1 tem valor médio $\bar{R}$ e variância $\left(1 - \frac{n}{N}\right) \frac{\sigma_R^2}{n}$.

Mas geralmente, $\bar{R}$ não é igual a R, e tem-se

$$\begin{aligned} \text{viés}(r_1) &= \bar{R} - R \\ &= -\frac{1}{X_T}\sum_{i=1}^{N} R_i(X_i - \bar{X}) \\ &= -\frac{(N-1)\sigma_{RX}}{X_T} \end{aligned} \tag{1.23}$$

onde σ_{RX} é a covariância entre R e X,

$$\sigma_{RX} = \frac{1}{N-1}\sum_{i=1}^{N}(R_i - \bar{R})(X_i - \bar{X}) = \frac{1}{N-1}\sum_{i=1}^{N} R_i(X_i - \bar{X})$$

Então, sabendo que o erro quadrático médio é a soma da variância com o quadrado do viés, tem-se que

$$\begin{aligned} EQM(r_1) &= Var(r_1) + (\text{viés}(r_1))^2 \\ &= \left(1 - \frac{n}{N}\right)\frac{\sigma_R^2}{n} + \frac{(N-1)^2\sigma_{RX}^2}{X_T^2} \end{aligned} \tag{1.24}$$

e um estimador centrado da covariância, σ_{RX} é

$$\frac{1}{n-1}\sum_{i=1}^{n} r_i(x_i - \bar{x}) = \frac{n}{n-1}(\bar{y} - \bar{x}\bar{r}) \tag{1.25}$$

Portanto, pode-se estimar o viés e o erro quadrado médio (EQM) de r_1 por

$$\widehat{\text{viés}(r_1)} = -\frac{(N-1)n(\bar{y} - r_1\bar{x})}{(n-1)X_T}$$

e

$$\widehat{EQM(r_1)} = (1-f)\frac{\sum_{i=1}^{n}(r_i - r_1)^2}{n} + \frac{(N-1)^2 n^2(\bar{y} - r_1\bar{x})^2}{(n-1)^2 X_T^2},$$

respectivamente, desde que o total X_T seja conhecido. Esta condição é, muitas vezes, satisfeita na prática.

Assim, se X_T for conhecido, pode corrigir-se r_1 com a estimativa do viés, obtendo-se o estimador modificado

$$r_1' = r_1 + \frac{(N-1)n(\bar{y} - r_1\bar{x})}{(n-1)X_T}$$

que é conhecido como estimador de Hartley-Ross.

Estimador r_2

Este estimador é mais utilizado que o abordado no ponto anterior. Embora seja enviesado e tenha distribuição assimétrica, em amostras grandes o viés é desprezável e a sua distribuição aproxima-se da distribuição normal, permitindo realizar inferências sobre R com base na distribuição normal de variância $Var(r_2)$.

Tal como em r_1, está-se perante a complicação de que tanto o numerador, $\bar{y}$, como o denominador, $\bar{x}$, apresentam uma variação aleatória. Vai-se começar mais uma vez com a determinação do viés. Note-se que, tendo em conta o desenvolvimento em série de Taylor em tordo de $\bar{X}$,

$$\begin{aligned} r_2 - R = \frac{\bar{y}}{\bar{x}} - R &= \frac{\bar{y} - R\bar{x}}{\bar{X}} \left(1 + \frac{\bar{x} - \bar{X}}{\bar{X}}\right)^{-1} \\ &= \frac{\bar{y} - R\bar{x}}{\bar{X}} \left[1 - \frac{\bar{x} - \bar{X}}{\bar{X}} + \left(\frac{\bar{x} - \bar{X}}{\bar{X}}\right)^2 - \dots\right] \end{aligned} \quad (1.26)$$

Como uma aproximação do viés pode-se considerar os dois primeiros termos da série e obter

$$E(r_2) - R = E\left(\frac{\bar{y} - R\bar{x}}{\bar{X}}\right) - \frac{1}{\bar{X}^2} E[(\bar{y} - R\bar{x})(\bar{x} - \bar{X})]$$

O termo principal é zero desde que $E(\bar{y} - R\bar{x}) = \bar{Y} - R\bar{X} = 0$. Assim,

$$E[\bar{y}(\bar{x} - \bar{X})] = Cov(\bar{y}, \bar{x}) = \frac{(1-f)\sigma_{YX}}{n} = \frac{(1-f)\rho_{YX}\sigma_Y\sigma_Y}{n}$$

onde ρ_{YX} é a correlação entre Y e X. Assim, uma aproximação para o viés é

$$\text{viés}(r_2) = E(r_2) - R \approx \frac{(1-f)}{n\bar{X}^2} \left(R\sigma_X^2 - \rho_{YX}\sigma_Y\sigma_X\right) \quad (1.27)$$

que será pequeno se ρ_{YX} não diferir muito de $R\frac{\sigma_X}{\sigma_Y}$. Isto equivale a dizer que a regressão de Y em X é linear e passa pela origem, isto é, que Y e X são aproximadamente proporcionais.

Para grandes amostras podem-se utilizar resultados assimptóticos e tem-se

$$E(r_2) \approx \frac{\bar{Y}}{\bar{X}} = \frac{Y_T}{X_T} = R$$

e

$$Var(r_2) \approx \frac{1-f}{n\bar{X}^2} \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - RX_i)^2$$

A variância de r_2 pode ser estimada por

$$\begin{aligned} s^2(r_2) &= \frac{1-f}{n\bar{x}^2} \frac{1}{N-1} \sum_{i=1}^{N} (y_i - r_2 x_i)^2 \\ &= \frac{1-f}{n(n-1)\bar{x}^2} \left\{ \sum_{i=1}^{n} y_i^2 - 2r_2 \sum_{i=1}^{n} y_i x_i + r_2^2 \sum_{i=1}^{n} x_i^2 \right\} \end{aligned}$$

Para grandes amostras, a distribuição de r_2 aproxima-se da distribuição normal, o que permite construir intervalos de confiança para R. Um intervalo de confiança a aproximadamente $100(1-\alpha)\%$ para R é dado por $r_2 \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right) s(r_2)$.

Exemplo 1.3.1: Conduziu-se um inquérito sobre o aumento do preço da comida recolhendo uma amostra aleatória simples de 48 produtos básicos de alimentação num hipermercado. Os preços desses 48 produtos alimentares foram registados em duas ocasiões diferentes, com um intervalo de 3 meses. Os preços registados pela primeira vez são designados por x_i e os da segunda vez por y_i. A razão das médias amostrais, $r_2 = \frac{\bar{y}}{\bar{x}}$ dá uma indicação da mudança do preço da alimentação durante os 3 meses em questão, já que se trata de uma estimativa da razão populacional R entre os preços médios nas duas ocasiões. Obtiveram-se os resultados:

$\bar{y} = 12.07, \qquad \bar{x} = 11.41;$

$\sum_{i=1}^{48} y_i^2 = 9270.6, \qquad \sum_{i=1}^{48} x_i^2 = 8431.7, \qquad \sum_{i=1}^{48} y_i x_i = 8564.1.$

A dimensão N da população (número de produtos alimentares distintos) é grande relativamente à dimensão da amostra, $n = 48$ e, portanto, pode-se ignorar a correcção de população finita, c.p.f. (1-f). Tem-se $r_2 = \frac{\bar{y}}{\bar{x}} = 1.06$, isto é, estima-se um aumento de 6% nos preços da alimentação durante os 3 meses do estudo.

A variância amostral (aproximada) de r_2 é

$$\frac{9270.6 - 2 * 1.06 * 8564.1 + (1.06)^2 * 8431.7}{48 * 47 * (11.41)^2} = (0.0447)^2$$

e um intervalo de confiança a 95% para R é

$$(1.06 \pm 1.96 * 0.0447) = (0.970; 1.145).$$

Com base neste intervalo de confiança aproximado, não é possível afirmar com firmeza que houve um aumento no preço médio da alimentação nos 3 meses de estudo. Note-se que a grande amplitude do intervalo de confiança reflecte a pequena dimensão da amostra.

1.3.2 Estimador da razão do total, Y_T, e da média, $\bar{Y}$

Suponha-se que se pretende estimar a despesa total dos municípios dum determinado país, num serviço em particular (saúdo ou educação, por exemplo) num determinado ano. Para tal, pode-se obter uma amostra aleatória simples de n municípios, registar as respectivas despesas e estimar Y_T por $y_T = N\bar{y}$. Mas, é evidente que vai haver uma grande diferença entre as quantidades gastas, em saúde ou educação, nos diferentes municípios, devido a várias razões (tais como a área do município, o número de habitantes, etc.). Seria desejável utilizar informação adicional sobre a estrutura da população de modo a obter um estimador mais eficiente do total Y_T do que y_T. Vejamos como utilizar essa informação adicional para construir um estimador de razão de Y_T (ou de $\bar{Y}$).

Suponha-se que Y_i representa a despesa do município i em saúde ou em educação, X_i é o número de habitantes desse município e, para os municípios da amostra registam-se simultaneamente o valor das duas variáveis, obtendo uma amostra aleatória simples bivariada de dimensão n: $(y_1, x_1), \ldots, (y_n, x_n)$.

O número total de habitantes do país, X_T, é usualmente conhecido quase correctamente (por exemplo, a partir da último censo da população). Também se conhece N, o número de municípios do país. Mas pode-se estimar X_T com base na amostra, utilizando o estimador $x_T = N\bar{x}$, em que $\bar{x}$ é a média da amostra aleatória simples. Analogamente, pode-se estimar a despesa total Y_T (a característica em que estamos interessados) por $y_T = N\bar{y}$. A estimativa x_T não tem interesse em si própria, já que conhecemos X_T, mas dá-nos a vantagem importante de através da sua comparação com X_T, podemos inferir informalmente da representatividade da amostra. Se x_T for muito menor que X_T então, em virtude da proporcionalidade aproximada entre Y_i e X_i, podemos concluir que y_T vai subestimar Y_T, se x_T for demasiado grande então y_T será provavelmente também demasiado grande. Se a relação de proporcionalidade fosse exacta, teríamos

$$Y_i = RX_i,\ i = 1, \ldots, N \tag{1.28}$$

em que R é a razão populacional, $R = \frac{Y_T}{X_T} = \frac{\bar{Y}}{\bar{X}}$. Assim,

$$Y_T = RX_T,$$

e poderíamos estimar Y_T substituindo R pelo seu estimador r_2, obtendo o estimador de razão de Y_T,

$$y_{TR} = r_2 X_T = \frac{X_T}{x_T} y_T \tag{1.29}$$

A partir daqui, vai-se utilizar como estimador de R apenas o estimador r_2 que designaremos simplesmente por r.

O estimador y_{TR} é designado por estimador de razão do total da população por amostragem aleatória simples. Note-se que aumenta ou diminui o estimador simples y_T através de um factor de condensação $\frac{X_T}{x_T}$. Se x_T for maior do que X_T, este factor é menor que 1 e o estimador y_T é reduzido, se x_T for menor que X_T, $\frac{X_T}{x_T}$ é maior do que 1 e o estimador y_T é aumentado.

No caso de proporcionalidade exacta, (1.28) serve apenas para motivar o estimador y_{TR}. Se a proporcionalidade exacta não se verificar, o que acontece na prática, o estimador y_{TR} ainda faz sentido se houver uma proporcionalidade aproximada entre a variável de interesse, Y, e a covariável X.

Se estivermos interessados em estimar a média da população, $\bar{Y}$, os mesmos argumentos levam a utilizar o estimador da razão da média da população,

$$\bar{y}_R = r\bar{X} = \frac{\bar{X}}{\bar{x}}\bar{y}. \tag{1.30}$$

Os estimadores de razão têm um atractivo óbvio, mas é necessário identificar as circunstâncias em que, com estes estimadores, se obtém um acréscimo significativo de eficiência relativamente aos estimadores y_T e $\bar{y}$, que usam a informação adicional da covariável.

Considere-se o estimador $\bar{y}_R$. Como r é assintoticamente centrado como estimador de R, $\bar{y}_R$ também é um estimador assintoticamente centrado de $\bar{Y}$.

A variância aproximada de $\bar{y}_R$, para grandes amostras é

$$\begin{aligned} Var(\bar{y}_R) &= \frac{1-f}{n}\sum_{i=1}^{N}\frac{(Y_i - RX_i)^2}{N-1} \\ &= \frac{1-f}{n}(\sigma_Y^2 - 2R\rho_{YX}\sigma_Y\sigma_X + R^2\sigma_X^2), \end{aligned} \tag{1.31}$$

onde $\rho_{YX} = \frac{\sigma_{YX}}{\sigma_Y\sigma_Y}$ é o coeficiente de correlação populacional entre Y e X. Se a proporcionalidade exacta (1.28) se verificasse, então $Var(\bar{y}_R)$ seria aproximadamente zero. Na prática tal não acontece, mas $Var(\bar{y}_R)$ é tanto menor quanto maior for a correlação positiva entre Y e X na população.

Os resultados para y_{TR} são análogos. y_{TR} é assintoticamente centrado e a sua variância, para grandes amostras, é

$$\begin{aligned} Var(y_{TR}) &= \frac{N^2(1-f)}{n}\sum_{i=1}^{N}\frac{(Y_i - RX_i)^2}{N-1} \\ &= \frac{N^2(1-f)}{n}(\sigma_Y^2 - 2R\rho_{YX}\sigma_Y\sigma_X + R^2\sigma_X^2), \end{aligned} \tag{1.32}$$

Mais uma vez é necessário estimar $Var(\bar{y}_R)$ e $Var(y_{RT})$ a partir da amostra e utiliza-se

$$\widehat{Var}(\bar{y}_R) = \frac{1-f}{n(n-1)}\left\{\sum_{i=1}^{n} y_i^2 - 2r\sum_{i=1}^{n} y_i x_i + r^2\sum_{i=1}^{n} x_i^2\right\} \tag{1.33}$$

e

$$\begin{aligned}\widehat{Var}(y_{TR}) &= \frac{N^2(1-f)}{n(n-1)}\left\{\sum_{i=1}^{n} y_i^2 - 2r\sum_{i=1}^{n} y_i x_i + r^2\sum_{i=1}^{n} x_i^2\right\} \\ &= N^2 Var(\bar{y}_R), \end{aligned} \tag{1.34}$$

respectivamente. Usando a normalidade assimptótica destes estimadores e as expressões da variância para grandes amostras, podemos construir intervalos de confiança aproximados para $\bar{Y}$ ou para Y_T da maneira usual. Uma regra prática razoável para utilizar a distribuição normal e a expressão aproximada para a variância dos estimadores é

- $n \geq 40$;
- $f = \frac{n}{N} \leq 0.25$;
- $C_Y = \frac{\sigma_Y}{\bar{Y}} \leq 0.1$;
- $C_X = \frac{\sigma_X}{\bar{X}} \leq 0.1$.

onde C_Y e C_X são chamados de coeficientes de variação populacionais para Y e X, respectivamente.

Quando os resultados para grandes amostras não são apropriados, o estabelecimento das propriedades de $\bar{y}_R$ e y_{TR} e a construção de intervalos de confiança para $\bar{Y}$ e Y_T são muito complicados e não estão totalmente estudados. Existem alguns resultados aproximados, mas não serão aqui abordados. Alguns destes resultados encontram-se sumariados em Couchran(1977, capítulo 6).

Já examinamos com algum detalhe as propriedades dos estimadores de razão da média ou do total de uma população, mas uma questão fundamental mantém-se. *Em que circunstâncias, se algumas, devemos preferir um estimador de razão da média ou do total ao de uma média ou do total de uma amostra aleatória simples?*. Isto é, quando é que $\bar{y}_R$ (ou y_{TR}) é mais eficiente do que $\bar{y}$ (ou y_T)?

A resposta vai depender do coeficiente de correlação populacional ρ_{YX} e dos coeficientes de variação populacionais C_Y e C_X. Temos que identificar as condições em que $Var(\bar{y}_R)$ é menor que $Var(\bar{y})$, isto é, em que o estimador

de razão é mais eficiente. Ora, da expressão de $Var(\bar{y})$ e da expressão aproximada de $Var(\bar{y}_R)$, conclui-se que

$$Var(\bar{y}_R) < Var(\bar{y}) \qquad \text{se} \qquad R^2\sigma_X^2 < 2R\rho_{YX}\sigma_Y\sigma_X \tag{1.35}$$

isto é, se

$$\rho_{YX} > \frac{1}{2}\frac{C_X}{C_Y}. \tag{1.36}$$

Portanto, não é certo que a utilização de $\bar{y}_R$ resulte num aumento da eficiência relativamente a $\bar{y}$. Assim, é necessário que ρ_{YX} seja suficientemente grande (na prática, temos que verificar o critério anterior utilizado as estimativas amostrais de C_Y, C_X e ρ_{YX}).

Mas note-se que mesmo sendo ρ_{YX} muito grande, nem sempre $\bar{y}_R$ (ou y_{TR}) é mais eficiente do que $\bar{y}$ (ou y_T). Se $C_X > 2C_Y$, a desigualdade acima nunca pode ser verificada, o que significa que, neste caso, o estimador $\bar{y}_R$ (ou y_{TR}) não pode ser mais eficiente do que $\bar{y}$ (ou y_T), mesmo que exista uma correlação positiva exacta entre Y e X.

Pode-se concluir que existem dois factores importantes que contribuem para o aumento de eficiência dos estimadores de razão:

- a variabilidade dos valores da variável auxiliar X não pode ser muito maior do que a de Y;
- o coeficiente de correlação ρ_{YX} tem que ser positivo e elevado.

No entanto, em muitas situações práticas estas condições são verificadas e os estimadores de razão constituem uma melhoria substancial relativamente a $\bar{y}$ ou y_T.

Resumindo, para utilizar estimadores de razão é necessário que:

(i) seja possível observar simultaneamente duas variáveis Y e X que sejam aproximadamente proporcionais (isto é, que tenham correlação positiva e elevada);

(ii) a variável auxiliar X não pode ter um coeficiente de variação muito maior do que o de Y;

(iii) a média populacional $\bar{X}$, ou o total X_T, têm que ser conhecidos.

A proporcionalidade em (i) implica que existe uma relação aproximadamente linear entre Y e X que passa pela origem. Se Y e X tiverem uma relação aproximadamente linear que não passe pela origem, é preferível utilizar um estimador alternativo, conhecido como estimador de regressão, que será abordado na secção seguinte.

1.3.3 Estimadores de regressão

O estimador de regressão é útil quando existe algum grau de linearidade entre Y e X que não passa pela origem. Este estimador pode usar-se em situações em que $\bar{X}$ é conhecido.

Uma relação exacta pode ser escrita da forma

$$Y_i = \bar{Y} + B(X_i - \bar{X}) \tag{1.37}$$

para todos os valores da população (Y_i, X_i) e para algum valor de B. Se tal relação fosse verdadeira, poderíamos determinar exactamente $\bar{Y}$ a partir de uma só observação (y, x) pois

$$\bar{Y} = y - B(x - \bar{X})$$

Mas na prática isto não acontece. Poder-se-ia considerar, antes, o modelo

$$Y_i = \bar{Y} + B(X_i - \bar{X}) + E_i, \qquad i = 1, \ldots, N \tag{1.38}$$

assumindo que $\bar{E} = 0$, os valores de E_i não estão correlacionados com os X_i, $i = 1, \ldots, N$ (isto é, $\sigma_{XE} = 0$) e $\sigma_E^2 << \sigma_Y^2$.

O modelo (1.38) é uma representação adequada de uma população em que a variação dos valores de Y se deve em parte a uma dependência linear dos valores correspondentes de X. Sob o modelo (1.38), tem-se

$$\sigma_Y^2 = B^2\sigma_X^2 + \sigma_E^2$$

e, o coeficiente de correlação é

$$\rho_{YX} = B\frac{\sigma_X}{\sigma_Y}$$

o que implica que $\sigma_E^2 = \sigma_Y^2(1-\rho_{YX}^2)$.

A partir de uma amostra aleatória simples $(y_1, x_1), \ldots, (y_n, x_n)$ e supondo que $\bar{X}$ é conhecido, podemos considerar o estimador de regressão linear de $\bar{Y}$ dado por

$$\bar{y}_L = \bar{y} + b(\bar{X} - \bar{x}) \tag{1.39}$$

e o estimador de regressão linear de Y_T

$$y_{TL} = N\bar{y}_L \tag{1.40}$$

O estimador $\bar{y}_L$ é um estimador centrado de $\bar{Y}$, uma vez que

$$E(\bar{y}_L) = E(\bar{y}) + B(\bar{X} - E(\bar{x})) = \bar{Y},$$

e

$$\begin{aligned} Var(\bar{y}_L) &= E[(\bar{y}_L - \bar{Y})^2] \\ &= E\{[(\bar{y} - \bar{Y}) - B(\bar{x} - \bar{X})]^2\} \\ &= Var(\bar{y}) + B^2 Var(\bar{x}) - 2BCov(\bar{y}, \bar{x}) \\ &= \frac{1-f}{n}(\sigma_Y^2 - 2B\sigma_{YX} + B^2\sigma_X^2) \\ &= \frac{1-f}{n}\sigma_Y^2(1-\rho_{YX}^2) \end{aligned} \tag{1.41}$$

Portanto, para o modelo (1.38) tem-se que $Var(\bar{y}_L) \leq Var(\bar{y})$ e a eficiência de $\bar{y}_L$ relativamente a $\bar{y}$ aumenta com ρ_{YX}^2.

O estimador $\bar{y}_L$ é centrado qualquer que seja a dimensão da amostra, e um estimador centrado de $Var(\bar{y}_L)$ é

$$\widehat{Var}(\bar{y}_L) = \frac{1-f}{n}(s_Y^2 - 2Bs_{YX} + B^2 s_X^2),$$

onde s_Y^2, s_X^2 e s_{YX} são os estimadores centrados usuais de σ_Y^2, σ_X^2 e σ_{YX}, respectivamente. Por exemplo,

$$s_{YX} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}). \tag{1.42}$$

Na prática, o valor exacto de B é desconhecido e o modelo (1.38) com a hipótese $\sigma_{XE} = 0$ não se verifica exactamente. O estudo do estimador $\bar{y}_L$ sob o modelo (1.38) serve, apenas, para motivar a utilização da família dos estimadores de regressão linear do tipo

$$\bar{y}_L = \bar{y} + b(\bar{X} - \bar{x}) \tag{1.43}$$

como um principio geral de estimação. Vamos estudar as propriedades de $\bar{y}_L$ em condições mais gerais de dependência entre Y e X. Devemos considerar duas possibilidades: o valor de b é prefixado, o valor de b é estimado a partir da amostra.

(a) b **prefixado**

Qualquer que seja o valor de b, $\bar{y}_L$ é estimador centrado de $\bar{Y}$ e

$$Var(\bar{y}_L) = \frac{1-f}{n}(\sigma_Y^2 - 2b\sigma_{YX} + b^2\sigma_X^2).$$

com um estimador centrado dado por

$$\widehat{Var}(\bar{y}_L) = \frac{1-f}{n}(s_Y^2 - 2bs_{YX} + b^2 s_X^2).$$

Uma questão que surge é a seguinte: se $\bar{y}_L$ é um estimador centrado para todos os valores de b, para que valor de b a variância é miníma? Ora, minimizar $Var(\bar{y}_L)$ é equivalente a minimizar $-2b\sigma_{YX} + b^2\sigma_X^2$ e o mínimo desta expressão é atingido se

$$2b\sigma_X^2 - 2\sigma_{YX} = 0 \Leftrightarrow b = b_0 = \frac{\sigma_{YX}}{\sigma_X^2} = \rho_{YX}\frac{\sigma_Y}{\sigma_X}$$

Portanto, o mínimo de $Var(\bar{y}_L)$ é $\frac{1-f}{n}\sigma_Y^2(1-\rho_{YX}^2)$ e o estimador

$$\bar{y}_L = \bar{y} + \rho_{YX}\frac{\sigma_Y}{\sigma_X}(\bar{X} - \bar{x})$$

é o estimador mais eficiente de $\bar{Y}$ da forma (1.43), independentemente de qualquer relação existente entre Y e X na população.

Mas, na prática b_0 é desconhecido e, portanto, o estimador óptimo é inacessível. Contudo, pode ser razoável estipular um certo valor para b, com base em estudos anteriores de natureza similar.

Neste caso, para averiguar como é que o estimador considerado se compara em eficiência com o estimador da forma (1.43), podemos considerar a eficiência relativa:

$$\frac{1-\rho_{YX}^2}{1-2b\rho_{YX}\frac{\sigma_X}{\sigma_Y}+b^2\frac{\sigma_X^2}{\sigma_Y^2}} = \left[1+\frac{\rho_{YX}^2(1-\frac{b}{b_0})^2}{1-\rho_{YX}^2}\right]^{-1} \tag{1.44}$$

que pode ser estimada, para grandes amostras, substituindo σ_Y, σ_X e σ_{YX} pelos seus estimadores s_Y, s_X e s_{YX}.

A expressão da eficiência relativa implica que uma escolha de b afastada do óptimo valor b_0 pode resultar numa grande perda de eficiência do estimador de regressão linear. A ineficiência relativa será maior em populações em que os valores de Y e X estejam altamente correlacionados. Se a correlação for modesta, a escolha de b é menos importante mas, por outro lado, o ganho potencial relativamente a $\bar{y}$ é muito menor.

(b) b **estimado**

Mesmo que não se tenha nenhuma base para atribuir um valor a b, o que acontece geralmente, temos que o estimar a partir dos dados. O Valor óptimo de b, $\rho_{YX}\frac{\sigma_Y}{\sigma_X}$, obtido anteriormente, sugere que se estime b através da correspondente expressão amostral

$$\begin{aligned}\widetilde{b} &= \frac{s_{YX}}{s_X^2} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n y_i x_i - \bar{x}\sum_{i=1}^n y_i - \bar{y}\sum_{i=1}^n x_i + n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}} \end{aligned} \tag{1.45}$$

O estimador de regressão da média de uma população tem a forma

$$\bar{y}_L = \bar{y} + \widetilde{b}(\bar{X} - \bar{x}). \tag{1.46}$$

As propriedades deste estimador são difíceis de determinar com exactidão, uma vez que se tem presente uma variável aleatória adicional, $\widetilde{b}$, que é a razão de duas estatísticas.

Para grandes amostras as propriedades de (1.46) são mais facilmente estudadas. Este estimador é aproximadamente centrado e

$$Var(\bar{y}_L) \approx \frac{1-f}{n}\sigma_Y^2(1 - \rho_{YX}^2) \tag{1.47}$$

Assim, para grandes amostras é preferível estimar b por $\widetilde{b}$ em vez de atribuir um valor a b.

A estimativa da variância é dada por

$$s^2(\bar{y}_L) = \frac{1-f}{n}(s_Y^2 - \widetilde{b}s_{YX}).$$

1.3.4 Comparação dos estimadores de razão e de regressão

Tendo em conta as expressões para $Var(\bar{y}_R)$ e $Var(\bar{y}_L)$ para grandes amostras, verifica-se que

$$\begin{aligned} Var(\bar{y}_R) - Var(\bar{y}_L) &\approx \frac{1-f}{n}(R^2\sigma_X^2 - 2R\rho_{YX}\sigma_Y\sigma_Y + \rho_{YX}^2\sigma_Y^2) \\ &= \frac{1-f}{n}(R\sigma_X - \rho_{YX}\sigma_Y)^2 \geq 0 \end{aligned} \tag{1.48}$$

e, consequentemente, para grandes amostras o estimador de regressão é pelo menos tão eficiente como o estimador de razão, sob todas as circunstâncias. De (1.48) pode-se verificar que a única situação em que o estimador da razão apresenta a mesma eficiência que o estimador de regressão é quando

$$R = \rho_{YX} \frac{\sigma_Y}{\sigma_X}, \tag{1.49}$$

isto é, se $R = b_0$.

Note-se que não é necessário admitir qualquer formulação explicita sobre uma possível relação linear entre Y e X para deduzir as propriedades de $\bar{y}$, $\bar{y}_R$ e $\bar{y}_L$ descritas acima. Assim, $\bar{y}_L$ é sempre mais eficiente do que $\bar{y}$, excepto no caso em que $\rho_{YX} = 0$ em que têm a mesma eficiência.

Finalmente, $\bar{y}_L$ é sempre mais eficiente do que $\bar{y}_R$, excepto no caso particular em que $\rho_{YX} = 0$ onde os estimadores têm a mesma eficiência.

1.4 Amostragem Aleatória Estratificada

Existem certos casos em que a população está naturalmente dividida em grupos. Outras vezes, por conveniência e facilidade de amostragem, divide-se a população em grupos. Em ambos os casos, diz-se que se trata de uma população estratificada.

Sob condições adequadas, a estratificação da população pode melhorar a eficiência dos estimadores das características da população.

Considere-se um exemplo numérico. Suponha-se que se tem uma população de 20 membros para os quais a variável Y toma os valores:

$$6 \quad 3 \quad 4 \quad 4 \quad 5 \quad 3 \quad 6 \quad 2 \quad 3 \quad 2 \quad 2 \quad 6 \quad 5 \quad 3 \quad 5 \quad 2 \quad 4 \quad 6 \quad 4 \quad 5$$

A média destes valores é $\bar{Y} = 4$ e a variância é $\sigma^2_Y = \frac{40}{19}$. Se extrairmos uma amostra aleatória simples de dimensão 5 e utilizarmos $\bar{y}$ para estimar $\bar{Y}$, tem-se que

$$Var(\bar{y}) = \frac{(1-f)\sigma^2}{n} = \frac{\left(1-\frac{5}{20}\right)\frac{40}{19}}{5} = 0.316$$

Evidentemente, consoante a amostra de dimensão 5 extraída, obtém-se uma estimativa diferente de $\bar{y}$, que varia entre 2.2 e 5.8. Mas se observarmos a estrutura da população e ordenarmos os valores de Y por ordem crescente

$$2 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 3 \quad 3 \quad 4 \quad 4 \quad 4 \quad 4 \quad 5 \quad 5 \quad 5 \quad 5 \quad 6 \quad 6 \quad 6 \quad 6$$

observa-se que a população é constituída por 5 grupos, em cada um dos quais o valor da Y é constante. Suponha-se que extraímos aleatoriamente um elemento de cada um destes grupos para obter uma amostra de dimensão 5. Com tal extracção vão-se obter, invariavelmente, os valores:

$$2 \quad 3 \quad 4 \quad 5 \quad 6$$

cuja média é 4. Portanto, deste modo o estimador não apresenta flutuações amostrais, isto é, a sua variância amostral é zero, e a estimativa é sempre igual à média $\bar{Y}$ da população.

Isto apenas acontece porque os grupos são tais que dentro de cada de cada um não existe variabilidade. Trata-se de um exemplo "extremo", mas permite ilustrar a possibilidade de reduzir a variância do estimador da média da população, dividindo a população em subgrupos relativamente homogéneos

(isto é, com reduzida variabilidade dentro de cada grupo) e extraindo aleatoriamente e sem reposição em certo número de membros de cada grupo para construir a amostra de dimensão n.

Vai-se agora ver como estimar as características da população em populações estratificadas e em que circunstâncias se obtêm melhores estimadores do que os estimadores baseados numa amostra aleatória simples da população não estratificada.

1.4.1 Amostragem aleatória (simples) estratificada

Suponha-se que se deseja estimar a média, $\bar{Y}$, de um conjunto de valores $Y_1, \ldots, Y_N$ numa população finita. Vamos assumir que a população está estratificada, isto é, que é constituída por k grupos disjuntos ou estratos de dimensões

$$N_1, \ldots N_k \qquad \left(\sum_{i=1}^{k} N_i = N\right)$$

com membros

$$Y_{ij} \qquad (i = 1, \ldots, k; j = 1, \ldots, N_i)$$

As médias e variâncias dos estratos (ou subpopulações) são designados por

$$\bar{Y}_1, \ldots, \bar{Y}_k$$

e

$$\sigma_1^2, \ldots, \sigma_k^2,$$

respectivamente.
A média $\bar{Y}$ e a variância σ^2 da população podem escrever-se da forma

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{Y}_i = \sum_{i=1}^{k} W_i \bar{Y}_i,$$

onde $W_i = \frac{N_i}{N}$ é o *peso* do estrato i, $i = 1, \ldots, k$, e

$$\begin{aligned}\sigma^2 &= \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y})^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y}_i+\bar{Y}_i-\bar{Y})^2 \\ &= \frac{1}{N-1}\left\{\sum_{i=1}^{k}(N_i-1)\sigma_i^2+\sum_{i=1}^{k}N_i(\bar{Y}_i-\bar{Y})^2\right\}. \qquad (1.50)\end{aligned}$$

Assume-se que uma amostra de dimensão n é escolhida por obtenção de uma amostra aleatória simples de cada estrato. As dimensões de cada estrato vão ser denotadas por $n_1,\ldots,n_k$ $\left(n=\sum_{i=1}^{k}n_i\right)$. A amostra aleatória simples proveniente do i-ésimo estrato tem como membros

$$y_{i1},\ldots,y_{in_i}, \qquad i=1,\ldots,k,$$

e a média e variância amostrais do i-ésimo estrato são dadas por

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

e

$$s_i^2=\frac{1}{n_i-1}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2.$$

Para cada estrato temos uma fracção de amostragem $f_i=\frac{n_i}{N_i}$, $i=1,\ldots,k$.

Este esquema de amostragem para a obtenção de uma amostra de dimensão total n do conjunto da população é chamado *amostragem aleatória (simples) estratificada*.

O estimador de $\bar{Y}$ usualmente utilizado é a *média amostral estratificada*:

$$\bar{y}_{st}=\sum_{i=1}^{k}W_i\bar{y}_i.$$

Note-se que se assume que se conhecem as dimensões dos estratos, N_i, e, portanto, os pesos dos estratos, $W_i=\frac{N_i}{N}$, $i=1,\ldots,k$.
A média amostral estratificada $\bar{y}_{st}$ não é, em geral, igual à média amostral:

$$\bar{y}' = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{y}_i = \sum_{i=1}^{k} \frac{n_i}{n}\bar{y}_i$$

da amostra aleatória estratificada. A igualdade apenas se verifica quando

$$\frac{n_i}{n} = \frac{N_i}{N}, \qquad 1 = 1, \ldots, k.$$

Isto implica que as fracções de amostragem $f_i = \frac{n_i}{N_i}$ são iguais em todos os estratos. Neste caso diz-se que as dimensões dos estratos, n_i, são escolhidas por *afectação proporcional*, já que os n_i são escolhidos por forma a serem proporcionais à dimensão dos estratos, isto é, $n_i = N_i \frac{n}{N}$, $i = 1, \ldots, k$.

Este procedimento pode simplificar a recolha dos dados e tem a vantagem do ponto de vista estatística, mas pressupõe que as dimensões dos estratos, N_i, são conhecidas. Se tal não acontecer, os pesos, W_i, têm que ser estimados e o estimador $\bar{y}_{st}$ passará a ser enviesado e perderá eficiência. No que se segue vamos supor que os N_i, $i = 1, \ldots, k$ são conhecidos.

O valor médio e variância de $\bar{y}_{st}$ são dados por

$$E(\bar{y}_{st}) = \sum_{i=1}^{k} W_i E(\bar{y}_i) = \sum_{i=1}^{k} W_i \bar{Y}_i = \bar{Y}$$

e

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2 Var(\bar{y}_i) = \sum_{i=1}^{k} W_i^2 (1 - f_i)\frac{\sigma_i^2}{n_i}, \tag{1.51}$$

já que $cov(\bar{y}_i, \bar{y}_j) = 0$ para $i \neq j$, isto é, as médias amostrais de estratos diferentes não são correlacionadas.

Note-se que $\bar{y}_{st}$ é um estimador centrado de $\bar{Y}$, e

$$E(\bar{y}') = \frac{1}{n}\sum_{i=1}^{k} n_i E(\bar{y}_i) = \sum_{i=1}^{k} \frac{n_i}{n}\bar{Y}_i$$

e, portanto, a média global da amostra estratificada só será centrada no caso de afectação proporcional. Convém ainda referir que $\bar{y}'$ não tem a mesma

variância que a média $\bar{y}$ de uma amostra aleatória simples de dimensão n extraída da globalidade da população. Esta diferença deve-se ao elemento de aleatoriedade da amostra aleatória estratificada, que é devido ao facto de números prefixados n_i de elementos da amostra terem de ser extraídos de cada um dos estratos definidos pela estratificação da população.

Devem-se considerar alguns casos especiais de (1.51)

(a) As fracções de amostragem, $f_i = \frac{n_i}{N_i}$, são desprezáveis,

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2 \frac{\sigma_i^2}{n}$$

.

(b) Afectação proporcional, $n_i = nW_i$, $f_i = f = \frac{n}{N}$,

$$Var(\bar{y}_{st}) = \frac{1-f}{n} \sum_{i=1}^{k} W_i \sigma_i^2$$

(c) Afectação proporcional e variâncias iguais nos estratos, $\sigma_i^2 = \sigma_W^2$, $i = 1, \ldots, k$,

$$Var(\bar{y}_{st}) = \frac{1-f}{n} \sigma_W^2.$$

Os resultados para a estimação do **total da população**, Y_T são análogos.

$$y_{T_{st}} = N\bar{y}_{st} = \sum_{i=1}^{k} N_i \bar{y}_i$$

é um estimador de Y_T com

$$Var(y_{T_{st}}) = \sum_{i=1}^{k} N_i^2 (1 - f_i) \frac{\sigma_i^2}{n_i}.$$

Na prática, as variâncias dentro de cada estrato, σ_i^2, não são conhecidas. Portanto, para aferir da precisão dos estimadores $\bar{y}_{st}$ e $y_{T_{st}}$ é necessário estimá-las. Como os estratos são apenas subpopulações e os valores da amostra pertencentes a cada um dos estratos constituem uma amostra aleatória simples desse estrato, os estimadores

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2, \ i = 1, \ldots, k$$

são estimadores centrados dos σ_i^2. Portanto, um estimador centrado de $Var(\bar{y}_{st})$ é dado por

$$\begin{aligned} s^2(\bar{y}_{st}) &= \sum_{i=1}^{k} W_i^2 (1 - f_i) \frac{s_i^2}{n_i} \\ &= \frac{1}{N^2} \sum_{i=1}^{k} N_i (N_i - n_i) \frac{s_i^2}{n_i} \end{aligned} \tag{1.52}$$

Naturalmente, é necessário que o número de estratos seja de pelo menos 2, isto é, $n_i \leq 2$, $i = 1, \ldots, k$.

Em algumas situações, as circunstâncias práticas sugerem que as variâncias dos estratos são todas iguais. Neste caso, é desejável combinar os dados relativos aos vários estratos para obter um estimador centrado da variância comum σ_W^2, esse estimador é da forma

$$s_W^2 = \frac{1}{N - k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Pode-se, agora, estimar a $Var(\bar{y}_{st})$ por

$$s^2(\bar{y}_{st}) = \frac{s_W^2}{N^2} \sum_{i=1}^{k} \frac{N_i (N_i - n_i)}{n_i}.$$

Nesta situação é conveniente usar afectação proporcional na extracção da amostra e, um estimador centrado de $Var(\bar{y}_{st})$ será apenas

$$s^2(\bar{y}_{st}) = (1 - f) \frac{s_W^2}{n}.$$

Assumindo, como habitualmente, uma distribuição aproximadamente normal para $\bar{y}_{st}$, podemos construir **intervalos de confiança** aproximados para $\bar{Y}$

ou Y_T. Assim, para um grau de confiança $(1-\alpha)100\%$, estes intervalos são dados por

$$\left[\bar{y}_{st} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right) s(\bar{y}_{st});\ \bar{y}_{st} + \Phi^{-1}\left(1-\frac{\alpha}{2}\right) s(\bar{y}_{st})\right] \tag{1.53}$$

e

$$\left[N\left(\bar{y}_{st} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right) s(\bar{y}_{st})\right);\ N\left(\bar{y}_{st} + \Phi^{-1}\left(1-\frac{\alpha}{2}\right) s(\bar{y}_{st})\right)\right], \tag{1.54}$$

respectivamente.

1.4.2 Comparação de $\bar{y}$ e $\bar{y}_{st}$

A estratificação da população pode, em certos casos, aumentar a eficiência da estimação de $\bar{Y}$ ou Y_T. Para examinar esta situação, vamos comparar os estimadores $\bar{y}$ e $\bar{y}_{st}$. estes estimadores são ambos centrados. Vejamos qual deles tem menor variância. Sabemos que

$$Var(\bar{y}) = (1-f)\frac{\sigma^2}{n}.$$

Para simplificar a comparação, vamos considerar que a amostra estratificada foi extraída com afectação proporcional. Então,

$$Var(\bar{y}_{st}) = (1-f)\sum_{i=1}^{k} W_i \sigma_i^2$$

e

$$Var(\bar{y}) - Var(\bar{y}_{st}) = \frac{1-f}{n}\left(\sigma^2 - \frac{1}{N}\sum_{i=1}^{k} N_i \sigma_i^2\right)$$

Mas, por (1.50),

$$\sigma^2 = \frac{1}{N-1}\left\{\sum_{i=1}^{k}(N_i-1)\sigma_i^2 + \sum_{i=1}^{k} N_i(\bar{Y}_i - \bar{Y})^2\right\}.$$

Se as dimensões dos estratos N_i forem suficientemente grandes, tem-se

$$\frac{N_i - 1}{N - 1} \approx \frac{N_i}{N - 1} \tag{1.55}$$

e

$$\sigma^2 \approx \frac{1}{N} \left\{ \sum_{i=1}^{k} N_i \sigma_i^2 + \sum_{i=1}^{k} N_i (\bar{Y}_i - \bar{Y})^2. \right\}$$

Então,

$$Var(\bar{y}) - Var(\bar{y}_{st}) \approx \frac{1 - f}{n} \sum_{i=1}^{k} W_i (\bar{Y}_i - \bar{Y})^2,$$

que é maior que zero, excepto no caso em que $\bar{Y}_i = \bar{Y}$, $i = 1, \ldots, k$.

Pode-se então concluir que a média amostral estratificada será sempre mais eficiente do que a média de uma amostra aleatória simples, $\bar{y}$, e a diferença é tanto maior quanto maior for a variação nas médias dos estratos, $\bar{Y}_i$.

Suponha-se, agora, que a hipótese (1.55) não é razoável, isto é, que as dimensões dos estratos são suficientemente grandes para que a aproximação (1.55) seja razoável. Nesse caso, obtém-se a expressão mais exacta

$$Var(\bar{y}) - Var(\bar{y}_{st}) = \frac{1 - f}{n(N - 1)} \left\{ \sum_{i=1}^{k} N_i (\bar{Y}_i - \bar{Y})^2 - \frac{1}{N} \sum_{i=1}^{k} (N - N_i) \sigma_i^2 \right\}, \tag{1.56}$$

que não é necessariamente positiva. Portanto, $\bar{y}_{st}$ não é necessariamente mais eficiente do que $\bar{y}$. $\bar{y}_{st}$ será mais eficiente do que $\bar{y}$ se

$$\sum_{i=1}^{k} N_i (\bar{Y}_i - \bar{Y})^2 > \frac{1}{N} \sum_{i=1}^{k} (N - N_i) \sigma_i^2 \tag{1.57}$$

Observe-se uma situação particular com uma interpretação mais simples desta condição. Suponha-se que todos os estratos têm a mesma variância, σ_W^2. Neste caso, a equação anterior pode ser escrita da forma

$$\frac{1}{k-1}\sum_{i=1}^{k} N_i(\bar{Y}_i - \bar{Y})^2 > \sigma_W^2. \tag{1.58}$$

Portanto, $\bar{y}_{st}$ será mais eficiente do que $\bar{y}$ se a variação entre as médias dos estratos for suficientemente grande comparada com a variação dentro de cada estrato.

Resumindo, pode concluir-se informalmente que quanto maior for a variabilidade nas médias dos estratos e quanto menor for a variabilidade dentro de cada um dos estratos, maior será o ganho potencial de utilizar $\bar{y}_{st}$ em vez de $\bar{y}$ para estimar $\bar{Y}$. O mesmo acontece para a estimação de Y_T.

1.4.3 Escolha óptima das dimensões das amostras dos estratos

Deve ser considerada de novo a questão da escolha da dimensão da amostra, n, de modo a satisfazer determinados requisitos de precisão ou de custos. Desde que diferentes estratos da população apresentem grau de variabilidade diferentes, deve-se, além da escolha de n, escolher também os valores da dimensão amostral de cada estrato, n_i.

No caso de custos de amostragem diferentes para diferentes estratos tem que se ter em consideração os factores de custo na determinação das dimensões dos diferentes estratos. O modelo de custos mais simples considera que existe um custo base c_0 de administração do inquérito por amostragem e que observações individuais do estrato i têm um custo adicional unitário de c_i, $i = 1, \ldots k$. Este custo é dado por:

$$C = c_0 + \sum_{i=1}^{k} c_i n_i \tag{1.59}$$

Este é o modelo que iremos adoptar, embora por vezes seja mais razoável substituir $\sum_{i=1}^{k} c_i n_i$ por $\sum_{i=1}^{k} c_i \sqrt{n_i}$, por exemplo.

Suponha-se que adoptamos o modelo de custos (1.59) e que pretendemos saber que afectação das dimensões das amostras dos estratos, $n_1, \ldots, n_k$ devemos adoptar para

(i) minimizar $Var(\bar{y}_{st})$, dado um custo total C.

(ii) minimizar o custo total C para um dado valor de $Var(\bar{y}_{st})$.

Vamos considerar os casos anteriores separadamente.

Variância mínima para um custo fixo

Temos que escolher valores para $n_1, \ldots, n_k$ que minimizem

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2 \frac{\sigma_i^2}{n_i} - \frac{1}{N} \sum_{i=1}^{k} W_i \sigma_i^2 \tag{1.60}$$

sujeito à restrição

$$\sum_{i=1}^{k} c_i n_i = C - c_0.$$

Utilizando o método dos multiplicadores de Lagrange vai-se obter a **afectação óptima para um custo total fixo** que é dada por

$$n_i = \frac{(C - c_0) W_i \frac{\sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{k} W_i \sigma_i \sqrt{c_i}}, \qquad i = 1, \ldots, k \tag{1.61}$$

e a dimensão total da amostra é dada por

$$n = \sum_{i=1}^{k} n_i = \frac{(C - c_0) \sum_{i=1}^{k} W_i \frac{\sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{k} W_i \sigma_i \sqrt{c_i}}. \tag{1.62}$$

Verifica-se que as dimensões amostrais dos estratos devem ser proporcionais às dimensões dos estratos, N_i, proporcionais aos desvios padrão dos estratos, σ_i, e inversamente proporcionais à raiz quadrada do custo de amostragem unitário em cada estrato. Estratos com grande variabilidade e baixo custo de amostragem unitário terão amostras maiores do que outros estratos.

No caso particular dos custos unitários c_i serem todos iguais tem-se que

$$C = c_0 + nc$$

e que c é o custo constante para os estratos. A afectação óptima é dada por:

$$n_i = \frac{W_i\sigma_i}{\sum_{i=1}^{k} W_i\sigma_i} n \tag{1.63}$$

com $n = \frac{C-c_0}{c}$. Esta afectação é conhecida como a **afectação de Neyman**. Pode ser equivalentemente vista como a afectação óptima para n fixo e ignorando a variação nos custos unitários para os vários estratos, no sentido de que, dado n, esta afectação minimiza $Var(\bar{y}_{st})$. Isto é, a minimização de $Var(\bar{y}_{st})$ sujeito a $\sum_{i=1}^{k} n_i = n$ conduz aos n_i da afectação de Neyman.

A variância mínima resultante da afectação de Neyman, isto é, para n fixo ignorando os custos de amostragem ou com um custo de amostragem fixo e custos unitários constantes, é dada por

$$Var_{min}(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k} W_i\sigma_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} W_i\sigma_i^2. \tag{1.64}$$

Custo mínimo para uma variância fixa

Suponhamos que em vez de colocar limite ao custo total, se fixa $Var(\bar{y}_{st})$. Pretendemos satisfazer a condição, para um valor prefixado V,

$$Var(\bar{y}_{st}) = V$$

para um custo total mínimo. Assim, pretende-se minimizar $\sum_{i=1}^{k} n_i c_i$ sujeito à restrição

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2 \frac{\sigma_i^2}{n_i} - \frac{1}{N}\sum_{i=1}^{k} W_i\sigma_i^2 = V.$$

Assim, temos que considerar

$$n_i = \frac{\sum_{i=1}^{k} W_i\sigma_i\sqrt{c_i}}{V + \frac{1}{N}\sum_{i=1}^{k} W_i\sigma_i^2} \frac{W_i\sigma_i}{\sqrt{c_i}}, \qquad i = 1, \ldots, k \tag{1.65}$$

e, como dimensão da amostra

$$n = \frac{\left(\sum_{i=1}^{k} W_i \sigma_i \sqrt{c_i}\right)\left(\sum_{i=1}^{k} W_i \frac{\sigma_i}{\sqrt{c_i}}\right)}{V + \frac{1}{N}\sum_{i=1}^{k} W_i \sigma_i^2}. \tag{1.66}$$

Mais uma vez, se os custos de amostragem unitários forem constantes ($c_i = c$, $i = 1, \ldots, k$) tem-se que, a afectação de Neyman é

$$n_i = \frac{W_i \sigma_i}{\sum_{i=1}^{k} W_i \sigma_i} \times n, \qquad i = 1, \ldots, k \tag{1.67}$$

e

$$n = \frac{\left(\sum_{i=1}^{k} W_i \sigma_i\right)^2}{V + \frac{1}{N}\sum_{i=1}^{k} W_i \sigma_i^2}. \tag{1.68}$$

Pode-se concluir que a afectação de Neyman é óptima para minimizar a dimensão total da amostra, já que isto equivale a minimizar o custo total, para uma dada variância de $\bar{y}_{st}$.

Vamos considerar mais uma situação. A afectação óptima pode não ser admissível, suponhamos que estamos a utilizar pesos de amostras prefixados $w_i = \frac{n_i}{n}$ para os diferentes estratos e pretendemos saber como determinar n de modo a obter $Var(\bar{y}_{st}) = V$ com V prefixado.

Dimensão da amostra necessária para obter uma certa $Var(\bar{y}_{st})$, para pesos amostrais dados

Pretende-se que $Var(\bar{y}_{st}) = V$. Realizando esta igualdade obtém-se

$$n = \frac{\sum_{i=1}^{k} W_i^2 \frac{\sigma_i^2}{w_i}}{V + \frac{1}{N}\sum_{i=1}^{k} W_i \sigma_i^2}.$$

Assim, uma primeira aproximação para a dimensão da amostra, n, pode ser dada por

$$n_0 = \frac{1}{V}\sum_{i=1}^{k} W_i^2 \frac{\sigma_i^2}{w_i},$$

ou, de uma forma mais precisa,

$$n = n_0 \left(1 + \frac{1}{NV} \sum_{i=1}^{k} W_i \sigma_i^2\right)^{-1}.$$

No caso particular de se ter afectação proporcional, $w_i = W_i$, e para a afectação de Neyman tem-se

$$n_0 = \frac{1}{V} \sum_{i=1}^{k} W_i \sigma_i^2, \qquad n = n_0 \left(1 + \frac{n_0}{N}\right)^{-1}$$

e

$$n_0 = \frac{1}{V} \left(\sum_{i=1}^{k} W_i \sigma_i\right)^2, \qquad n = n_0 \left(1 + \frac{1}{NV} \sum_{i=1}^{k} W_i \sigma_i^2\right)^{-1},$$

respectivamente.

1.4.4 Comparação da afectação proporcional e da afectação óptima

Uma questão que surge naturalmente é em que medida é que a afectação óptima é melhor que a afectação proporcional? A afectação proporcional não requer o conhecimento das variâncias dos estratos ou dos custos de amostragem. Vejamos qual é o ganho potencial em usar a afectação óptima em vez da afectação proporcional.

No que se segue apenas vai ser considerado um caso, a comparação da afectação proporcional com a afectação de Neyman (óptima para custos de amostragem constantes em cada estrato).

Vamos denotar a $Var(\bar{y}_{st})$ por V_P e V_N para a afectação proporcional e afectação de Neyman, respectivamente. Tem que se verificar a desigualdade $V_P \geq V_N$. Assim,

$$
\begin{aligned}
V_P - V_N &= \frac{1-f}{n}\sum_{i=1}^{k} W_i\sigma_i^2 - \left\{\frac{1}{n}\left(\sum_{i=1}^{k} W_i\sigma_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} W_i\sigma_i^2\right\} \\
&= \frac{1}{n}\left\{\sum_{i=1}^{k} W_i\sigma_i^2 - \left(\sum_{i=1}^{k} W_i\sigma_i\right)^2\right\} \\
&= \frac{1}{n}\sum_{i=1}^{k} W_i(\sigma_i - \bar{\sigma})^2 > 0
\end{aligned}
$$

com $\bar{\sigma} = \sum_{i=1}^{k} W_i\sigma_i$.

Pode-se concluir que quanto maior for a variabilidade das variâncias dos estratos, maior é a vantagem relativa da afectação óptima.

Vamos agora comparar $Var(\bar{y})$, em que $\bar{y}$ é a média de uma amostra aleatória simples, com $Var(\bar{y}_{st})$ supondo afectação de Neyman das dimensões das amostras dos estratos. Denotando $Var(\bar{y})$ por V e $Var(\bar{y}_{st})$ por V_N, tem-se que

$$V - V_N > 0$$

se

$$\left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2 - \frac{\bar{\sigma}^2}{n} + \frac{1}{N}\sum_{i=1}^{k} W_i\sigma_i^2 > 0.$$

Se supusermos que os N_i são suficientemente grandes de modo a que $\frac{N_i-1}{N-1} \approx \frac{N}{N-1}$ e $\frac{N}{N-1} \approx 1$ tem-se que

$$V - V_N > 0$$

se

$$\frac{1}{n}\sum_{i=1}^{k} W_i(\sigma_i - \bar{\sigma})^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i(\bar{Y}_i - \bar{Y})^2 > 0.$$

Portanto, verifica-se que existe um ganho de eficiência quando se utiliza uma amostra aleatória estratificada com afectação de Neyman, excepto no caso limite em que todos os $\bar{Y}_i$ são iguais e todas as variâncias dos estratos são iguais. A eficiência será tanto maior quanto maior for a variabilidade nas médias dos estratos ou nas variâncias dos estratos.