UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

PROBLEMAS MAL RESOLVIDOS EM AMOSTRAGEM

Não–Resposta e Formulação Tendenciosa de Questões

ANTÓNIO ALBERTO OLIVEIRA

Tese orientada pelo Professor Doutor Dinis Pestana
e pela Professora Doutora Fernanda Diamantino

MESTRADO EM ESTATÍSTICA

2011

Ao Rúben

Agradecimentos

Quero expressar a minha gratidão aos Professores Dinis Pestana e Maria Fernanda Diamantino, que pacientemente discutiram comigo o meu trabalho, e me guiaram na forma de o melhorar e atingir os objectivos que me tinha proposto.

Agradeço também ao Professor Fernando Sequeira, com quem estabeleci uma sólida relação de amizade, e que tanto me ajudou a superar ignorância e dificuldades em tantas cadeiras.

A minha relação com a Maria Isabel Almeida foi durante um largo período um incentivo, que quero reconhecer.

O Luís Miguel Onofre, apesar de estar a passar por um período de dificuldades familiares devido à saúde periclitante da sua mãe Maria Luísa, foi um apoio nobre e desinteressado, que muito me ajudou num período em que também a minha vida familiar atravessava uma situação difícil.

E o Gonçalo Silva foi o colega da FCUL com quem criei uma relação de amizade mais constante, que tanto ajuda nas agruras da vida profissional e de estudo.

Todos aqueles que me ajudaram e apoiaram, e que não cabe aqui nomear individualmente, sabem que os aprecio e estimo, e lhes agradeço.

Sumário

A Estatística transforma informação em conhecimento. É por isso essencial que a informação seja válida e suficiente. Um plano amostral adequado é o garante de que se obtém a informação relevante, e a determinação da dimensão que a amostra deve ter para que as estimativas tenham a precisão desejada é um problema central da Amostragem.

Usamos a amostragem aleatŕia aleatória simples sem reposição como exemplo condutor, complementando com uma exposição breve da teoria de Horvitz–Thompson. Infelizmente, a prática mostra que o fenómeno de não-resposta é incontornável, e a sua correcção com imputação múltipla não é decerto consensual. Inspirando-nos em Aleixo *et al.* (2007) e Diamantino (2008), abordamos o problema da não resposta quando a taxa de respostas é tão baixa que se pode admitir uma situação de rarefação.

Num mundo perfeito seria então de admitir que a sondagem permitiria conhecer, com elevado grau de confiança, os parâmetros de interesse da população. Mas com perguntas tendenciosas, se calhar "9 em cada 10 estrelas de cinema usam LUX". Apresentamos um inquérito simples sobre hábitos alimentares, em duas versões — enunciados neutros e enunciados tendenciosos — para ilustrar a perversão resultante de influenciar os inquiridos.

Abstract

Statistics is a tool for building up knowledge from information. It is therefore of the utmost importance that information is reliable and enough to attain the accuracy needed. An appropriate sampling design guarantees that the information is relevant, and one of the foremost problems of Sampling Theory is the computation of the sample size needed to obtain accurate parameter estimates, i.e., within a fixed error bound, at a predetermined confidence level.

We use simple random sampling without replacement as an illustrative exemple, discussing the Horvitz-Thompson theory as a general framework. The practice of surveys shows that non-response is almost unavoidable, and the correction of the ensuing bias using multiple imputation isn't widely accepted in its actual development stage. We tackle the problem of non-response when the response rate is so low that Rényi's (1956) rarefation seems to be an appropriate framework; our development follos closely Aleixo *et al.* (2007) and Diamantino (2008).

In a perfect world it would seem wise to believe in survey reports. But loaded questions can easily lead to marketing dubious statements as "9 out of 10 movie stars use LUX". We developed a simple questionnaire on feeding habitudes, in two different versions — one of them with straightforward unloaded questions, the other one with loaded questions indicating the wise and convenient answers — in order to exhibit the perversions and bias when the formulation of the questionnaire influences the answers of the respondents.

Conteúdo

Capítulo 1

Introdução

O desenvolvimento da Estatística mostrou que deve haver uma intencionalidade na obtenção dos dados "ao acaso", uma expressão técnica que teremos ocasião de discutir e precisar; são esses os dados representativos da população, capazes de fornecer estimativas sem viés dos parâmetros, com a precisão desejada.

No capítulo de abertura apresentamos uma panorâmica dos conceitos e técnicas de amostragem que nos pareceram mais relevantes. Existem excelentes livros elementares de amostragem, cuidando dos aspectos matemáticos, como por exemplo Barnett (1991) ou Thompson (1992), pelo que privilegiamos a exposição das ideias, remetendo para aqueles livros, ou para o clássico Cochran (1977) quem deseje mergulhar em textos matematicamente mais sofisticados. Os livros de Fink, de Bourque e de Oishi que constituem o *Survey Toolkit* (2003) publicado pela Sage são um auxiliar precioso para quem quiser conhecer o essencial das questões práticas de amostragem.

A *amostragem aleatória simples sem reposição* é a matriz da recolha de dados ao acaso. Depende de um arrolamento completo da população de que se pretende extrair a amostra, que nem sempre existe — caso em que a *amostragem sistemática* é uma alternativa natural. No caso de a população poder ser particionada em "estratos", pode proceder-se a amostragem aleatória simples sem reposição nesses estratos, e usar o teorema da probabilidade total para estimar os parâmetros de interesse. Claro que isso só tem vantagem se o equilíbrio da variância dentro dos estratos e entre os estratos permitir obter a mesma eficiência a mais baixo custo. Nesta *amostragem estratificada* põe-se naturalmente a questão da dimensão das subamostras; a solução elementar é todas proporcionais à dimensão das subpopulações, mas a solução inteligente é dimensioná-las tendo em atenção não só a dimensão dos estratos e o custo por unidade amostral em cada estrato, mas também a variabilidade em cada um desses estratos.

A amostragem estratificada faz sentido quando o número de subpopulações é baixo: nesse caso os estratos são *recenseados*, e procede-se a amostragem dentro de cada estrato. No caso de a partição da população ter muitos subconjuntos, cada qual

com um número escasso de elementos, procede-se ao contrário: escolhem-se alguns desses grupos por amostragem, e procede-se ao recenseamento populacional dentro de cada um deles (pode haver várias etapas de amostragem até chegar aos grupos que vão ser recenseados: por exemplo, na primeira etapa escolher por amostragem alguns bairros de Lisboa, na segunda etapa escolher por amostragem alguns quarteirões desses bairros, na terceira etapa escolher por amostragem alguns prédios desses quarteirões, e finalmente recensear todos os animais domésticos que existem em cada um dos prédios seleccionados). É uma estratégia amostral que naturalmente conduz ao mais baixo preço de informação por unidade amostral, muitas vezes excessiva relativamente ao que se pretende. Note-se que do mais baixo preço por unidade amostral não decorre que seja a estratégia amostral que conduz aos preços mais baixos.

As estratégias amostrais clássicas apostam no *plano de amostragem*, que procuram produzir estimadores centrados dos parâmetros em que estamos interessados. Neles a incerteza deriva apenas da probabilidade de inclusão de cada elemento da população na amostra. Nesse sentido, a amostragem em populações finitas afasta-se da corrente principal do pensamento estatístico.

Há por outro lado toda uma vertente de amostragem que em vez de ser *design based* é *model based*, nomeadamente usando-se um modelo de regressão linear, em geral incompleto, que leva aos estimadores de razão e aos estimadores de diferença, estes últimos particularmente úteis em auditorias, todos eles imprescindíveis quando há razões plausíveis para observar uma imagem do que se pretende, desde que se possa admitir que essa imagem varia como função aproximadamente linear do que é nosso objectivo conhecer. Também se usam modernamente estratégias amostrais que decorrem de se admitir um modelo probabilista para a população — por exemplo, havendo razões para modelar a população com uma Poisson, a amostragem por *quadrats*, ou os modelos para o decaímento da capacidade de observação com distância em amostragem por distâncias.

No **Capítulo 2** faremos uma exposição um pouco mais detalhada das estratégias amostrais mais comuns, usando a *aassr* — amostragem aleatória simples sem reposição como exemplo demonstrativo, complementada com uma breve referência ao estimador de Horwitz–Thompson (1952) como um enquadramento elegante para as questões fundamentais de amostragem.

A determinação da dimensão n que uma amostra deve ter para produzir estimativas do parâmetro de interesse com a acribia desejada é, porventura, o mais relevante contributo da Amostragem para o avanço da Ciência. No entanto, esse notável conjunto de resultados parte do pressuposto de que "tudo vai correr bem", isto é que fixado esse n se podem seleccionar n unidades amostrais ao acaso no rol da população. Ora na prática é muito comum uma parte das unidades amostrais escolhidas "escaparem" da amostra, por exemplo recusando-se a serem observadas ou escusando-se a responder a questões que são o objecto da sondagem. A *não resposta*

é um obstáculo importante ao uso efectivo da metodologia de amostragem; como todos os obstáculos, leva a novos desenvolvimentos, havendo por exemplo um forte investimento na teoria da imputação múltipla (Singh, 2003, Cap. 12, e em particular pp. 1021–1025).

Outra possibilidade é recalcular a dimensão da amostra tendo em linha de conta a estimativa da taxa de resposta. Inspiramo-nos ainda num trabalho de Aleixo *et al.* (2007) e no desenvolvimento que teve na dissertação de Diamantino (2008) para tratar de formas alternativas de contornar o problema da não-resposta. A ideia de base é ampliar a dimensão da amostra que se deve recolher para atingir a precisão desejada de forma mais sofisticada do que a simples proporcionalidade intuitiva em geral usada. Para tal, recorremos a resultados convergentes sobre a rarefação de Rényi (1956) e as somas aleatórias geométricas de Kovalenko (1965). Uma parte substancial dos resultados decorre do uso do método delta (Pestana e Velosa, 2010, pp. 1046–1048), que encontrámos exposto com mais detalhe em Martins (2008) e em Diamantino (2008). Consultámos também, sem exaustividade, Chandra (1999), que contém informação detalhada.

No **Capítulo 3** referimos outros erros alheios à amostragem, para além da não-resposta (que muitos consideram ser o protagonista entre eles), inspirando-nos em Aleixo *et al.* (2011) e textos aí citados, nomeadamente os verbetes da *International Encyclopedea of Statistical Sciences* de Lovric (2011) a que tivemos acesso.

De facto, o "nem tudo correr bem", pode advir de, inadvertidamente ou com dolo, as questões serem apresentadas de forma viciada, induzindo as respostas do entrevistado, que de um modo geral quer dar de si uma imagem favorável; indicar que uma das respostas possíveis é "inteligente" ou "correcta" leva os indivíduos, sobretudo os de *status* social mais fragilizado, a darem respostas menos genuínas, e enviesa as estimativas — o que pode ser muito interessante para os fazedores de imagem da área de *marketing*, mas é cientificamente reprovável.

Neste Capítulo 3 apresentamos um questionário muito simples (mas uma simplicidade que é fruto de muito esforço de depuração) sobre hábitos alimentares de alunos de uma escola secundária e de alunos da FCUL. De facto, há dois questionários, que são uma importante variante um do outro: enquanto num deles as questões são apresentadas com distanciamento e isenção, no outro são acompanhadas por informações tendentes a influenciar o inquirido e viciar os resultados da sondagem. Fazemos um breve relatório comparativo do que se infere com as respostas a cada uma das variantes. Também no que se refere a este capítulo a dissertação de doutoramento de Diamantino (2008), em que se apresenta um inquérito para avaliar dificuldades de vida e de adaptação de estudantes dos PALOPs na Universidade de Lisboa, foi inspiradora.

Capítulo 2

Conceitos Gerais Sobre Amostragem

Em Estatística, amostra é um subconjunto de elementos extraídos de um conjunto chamado População. Hoje em dia é cada vez mais necessário quantificar a informação sobre os aspectos da vida humana. De facto, tenta-se descrever os fenómenos que nos rodeiam através de métodos estatísticos. Os métodos para a recolha, tratamento, apresentação e interpretação de dados estatísticos são cada vez mais exigentes face à necessidade das sociedades actuais. A Comunicação social tem contribuído para a divulgação da estatística relatando números de casos que fazem o dia-a-dia, como por exemplo, a subida ou descida da taxa de desemprego, a percentagem de acidentes de aviação, entre outros. Para satisfazer a necessidade de informação reduzindo os custos recolhe-se apenas uma parte da população, isto é, uma amostra, que posteriormente será analisada e fundamentada para o estudo em causa. Em suma, 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.

2.1 O Plano Amostral

O Plano amostral consiste em seleccionar os elementos a partir dos quais se recolhem os dados necessários para o estudo em questão. Churchill (1983) afirma que a possibilidade de realizar inferência da amostra para toda população depende do método de selecção dos elementos. Os processos requeridos para recolher a amostra iniciam-se com a clara definição da população alvo e posteriormente com a identificação da população a inquirir, constituindo um elo de ligação entre as decisões em estudo. Conceber e levar à pratica um estudo é um processo complexo envolvendo diversas fases interdependentes. Os pontos seguintes realçam as componentes essenciais deste Plano Amostral:

1. A primeira componente é decidir que tipo de população alvo se pretende para o estudo em causa. Assim, sendo é necessário em primeira mão definir um conjunto de elementos para os quais se pretende obter informação. É de referir que a população em estudo é constituída por unidades amostrais. Uma unidade do estudo, por sua vez, pode conter pessoas, famílias, instituições, sejam elas de carácter religioso, financeiro ou educacional. A adopção de uma determinada unidade amostral deve-se ao objectivo do estudo. Na delineação do universo são utilizadas as variáveis geográficas, demográficas e temporais, que ao mesmo tempo são comummente utilizadas para delimitar o próprio universo estatístico.

2. A segunda componente é identificar qual o tipo de informação que se pretende obter relativamente à população, sendo necessário conhecer as características associadas aos vários processos de recolha e análise da informação que se julga necessária. A realização de um recenseamento é uma das várias formas de recolher informação, a reflexão nas desvantagens deste processo, raramente conduz à sua adopção. Um método alternativo é a realização de sondagens, em que, ao invés de inquirir todos os indivíduos da população alvo, escolhe-se apenas um subconjunto dessa população. Deste subconjunto é possível inferir conclusões efectuadas através de técnicas de inferência estatística resultando uma generalização para toda a população. Por isso, a utilização da sondagem é hoje, em vários domínios, a resposta ao conhecimento de uma população tomando por base a amostra. A sondagem é uma listagem dos elementos da qual se vai seleccionar a amostra (Churchill, 1983). É de salientar a dificuldade residente na constituição destas listas, pois em muitos casos é impossível fazer coincidir a população alvo com a população a inquirir, ou seja, o universo que se pretende estudar e o universo que é efectivamente estudado nem sempre coincidem. Mesmo assim, a sondagem tem vantagens relativamente ao recenseamento, nomeadamente o custo, entre outros.

   Deve ainda referir-se que para inquirir uma população seriam necessários muitos mais agentes do que para inquirir uma amostra representativa dessa população. Há um risco bem conhecido de a partir de um certo número se recrutarem agentes de baixa qualidade, que comprometem a qualidade dos dados que recolhem (chegando mesmo a inventá-los). É uma razão importante para preferir a recolha de dados por amostragem.

3. A terceira componente consiste em delinear uma técnica amostral. Depois da identificação dos dados, surge uma questão, como deverão ser recolhidos e qual o instrumento a utilizar (questionário estruturado, por exemplo), e o passo seguinte, consiste em definir um processo de amostragem adequado. Os procedimentos de amostragem agrupam-se em duas categorias: amostragem aleatória e amostragem não aleatória. A amostragem aleatória ou probabilística baseia-se no facto de qualquer elemento da população poder ser

seleccionado, bastando para isso, conhecer todos os inquiríveis. Em relação às amostras não aleatórias não é possível calcular a probabilidade de um indivíduo ser seleccionado — o que tem como consequência que depois não será possível avaliar a variância dos estimadores que vão ser usados —, pois a sua escolha baseia-se quase sempre numa triagem arbitrária, de conveniência ou decorrente de preconceitos do investigador.

4. A quarta componente refere-se à determinação da dimensão da amostra. O cálculo da dimensão de uma amostra é muito útil se queremos partir para um estudo estatístico com alguma confiança sobre a possibilidade de, no futuro, extrapolar os resultados para a população, ou seja, a dimensão da amostra está relacionada com a precisão dos intervalos de confiança. É evidente que um aumento da dimensão da amostra conduzirá a um aumento da precisão, contudo os custos, custos estes que estão associados em termos de recursos humanos e simultaneamente compartilhado com a gestão de tempo, não favorecem a opção de uma amostra demasiado grande. Uma amostra demasiado pequena produzirá uma estimação de precisão inadequada. As condições de opção pela precisão desejada terão que ser pré-estabelecidas inicialmente, ou seja, o gasto máximo que podemos realizar, o intervalo de tempo para a realização do estudo, saber como equilibrar as necessidades em relação às várias características da população que estejam a ser estimadas (características de interesse), contornar por estimação o desconhecimento de alguns parâmetros da população (nomeadamente a variância de estimadores) que podem afectar a precisão dos estimadores.

2.2 Métodos de Amostragem

A teoria da amostragem estuda as relações existentes entre uma população e as amostras extraídas dessa população. 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 conhecimento científico da realidade, onde outros processos ou métodos alternativos, por razões diversas, não se mostram adequados ou até mesmo possíveis. Existem dois grandes grupos, que se destacam da seguinte forma:

2.2.1 Amostragem Não Aleatória

Os métodos de amostragem não aleatória são métodos que assentam numa base pragmática ou humanamente intuitiva com um critério objectivo de um estudo rápido e com menores custos. Neste processo não aleatório há unidades do universo

que são excluídas, ou seja, não têm possibilidade de ser escolhidas, tornando-se um dos principais factores inconvenientes. Esta subjectividade impossibilita determinar a probabilidade de um determinado elemento ser escolhido e requer uma avaliação da representatividade da amostra, factores, esses que pesam na credibilidade e grau de precisão dos resultados. Se por um lado as amostras não aleatórias permitem obter informação com custos mais reduzidos, mais rapidamente e com menores necessidades de pessoal, têm o inconveniente de não se saber com que grau de confiança são as conclusões obtidas generalizáveis à população. Isto não é motivo para as colocar de parte. Salientam-se os vários tipos de amostras não aleatórias:

Amostragem Intencional

A amostra intencional é composta por unidades seleccionadas pelo investigador com o objectivo de representar características típicas da população, sendo portanto delineada para atingir um objectivo pré-concebido pelo investigador. Um exemplo deste tipo de amostra é a escolha, em tempo de eleições legislativas, de localidades "representativas", sendo neste caso a representatividade assegurada pela escolha das localidades que no passado têm dado como vencedor o partido realmente eleito. Como a base de sondagem é desconhecida e o critério de selecção dos elementos da amostra não é objectivo, existem, nos resultados fornecidos por este tipo de amostra, enviesamentos consideráveis e difíceis de avaliar.

Amostragem *Snowball*

Esta técnica de amostragem requer por parte do investigador a capacidade para localizar de início um grupo de indivíduos que tenham as características desejadas (Churchill, 1983). É uma forma de amostra intencional 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 aumentado como uma bola de neve à medida que os respondentes identificam outros potenciais respondentes. É muito utilizada quando se pretende estudar pequenas populações muito específicas, como por exemplo os deficientes motores, os "sem abrigo", no entanto pode originar em resultados enviesados uma vez que as pessoas tendem a indicar o nome de pessoas íntimas ou amigos (com comportamentos e pensamentos similares).

Amostragem por Conveniência

É uma técnica de amostragem em que os elementos são escolhidos por conveniência ou por facilidade. Um exemplo deste tipo de amostragem é o caso em que os espectadores de um determinado programa são convidados a responder a um questionário. Outro exemplo, sendo o objectivo fazer um inquérito a pessoas com baixos rendimentos, escolhem-se as que costumam ter apoio psicológico numa dada instituição,

porventura com base na forma como costumam pagar para o obter. Para responder a esta questão é seleccionado convenientemente um bairro com moradores apresentando estas características. Há uma forte possibilidade de estas amostras serem enviesadas, pois os inquiridos auto-seleccionam-se, respondendo aqueles que têm especial interesse pelo assunto em causa. Salienta-se que as amostras obtidas desta forma não são representativas da população.

Amostragem por Quotas

Neste tipo de amostragem, a amostra é obtida dividindo a população em grupos ou estratos (utilizando o mesmo princípio da amostragem estratificada), sendo seleccionado um certo número (quota) de elementos de cada grupo de modo não aleatório, isto é, cabe ao investigador decidir quem fará parte da amostra. A amostra por quotas é uma amostra estratificada não aleatória, ou seja, uma amostra que é obtida dividindo a população em categorias e seleccionando um certo número (uma quota) de elementos, de modo não aleatório, de cada categoria. Esta estratégia de amostragem é particularmente útil na situação de inexistência de uma listagem da população.

A talhe de foice, descrevemos um procedimento muito usado para evitar que o entrevistador escolha de um modo completamente subjectivo os entrevistados, o método dos itinerários aleatórios: é utilizado dentro de localidades pré-estabelecidas e serve para orientar o entrevistador na selecção dos respondentes, quando as entrevistas são feitas na rua ou de porta a porta. Neste tipo de amostras, cada elemento da população tem uma certa probabilidade conhecida de ser incluído na amostra sendo possível avaliar a representatividade da amostra e predizer o grau de precisão dos resultados obtidos. Vantagens deste processo: rapidez com que pode ser efectuado, e o baixo custo por unidade amostral.

2.2.2 Amostragem Aleatória

Os métodos de amostragem aleatória são caracterizados por todos os elementos da população poderem ser seleccionados de acordo com uma probabilidade pré-definida com vista a avaliar objectivamente a precisão das estimativas das propriedades da população obtidas a partir da amostra. Há estudos em que se seleccionam primeiramente lares e depois dentro do lar selecciona-se aleatoriamente um membro para ser entrevistado. A probabilidade deste indivíduo ser escolhido depende do tamanho da família, que só é conhecido depois do contacto com o lar (Frankel e Frankel, 1974). Na amostragem aleatória os critérios de selecção dos elementos estão rigorosamente definidos, não permitindo que a subjectividade ou livre arbítrio do entrevistador ou do investigador intervenham na escolha dos elementos. 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, o que acontece (mesmo quando o objectivo não é esse) sempre que usa a opinião e a experiência para escolher as amostras. A amostragem aleatória inclui a possibilidade de matematicamente determinar a dimensão da amostra em função da precisão e do grau de confiança desejados para os resultados. Salienta-se que a amostragem aleatória não se rodeia só de vantagens. Existem sérios obstáculos que em muitos casos a tornam impraticável de tal forma que se tem que recorrer a outros métodos. Uma das principais dificuldades é a obtenção de uma lista 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. Outro factor a ter em conta são as não–respostas obtidas, este tipo de dificuldade origina um aumento de custos, pois implica a necessidade de contactar várias vezes o mesmo inquirido ou optar por outro inquirido. A amostragem aleatória é, sem dúvida, o processo mais caro, mas os custos são compensados pela fiabilidade dos resultados obtidos.

No caso de uma população finita com N elementos, $\{Y_1, Y_2, \ldots, Y_N\}$, a média populacional é definida como usualmente $\overline{Y} = \dfrac{1}{N}\displaystyle\sum_{k=1}^{N} Y_K$, mas é conveniente definir a variância como

$$\sigma^2 = \frac{1}{N-1}\sum_{k=1}^{N}\left(Y_k - \overline{Y}\right)^2.$$

Amostragem Aleatória Simples

Uma amostra aleatória simples (sem reposição) — que abreviadamente designaremos por *aassr* — de k elementos extraída de uma população de N elementos é qualquer das $\binom{N}{k}$ amostras possíveis, com igual probabilidade, $\dfrac{1}{\binom{N}{k}}$ de ser seleccionada (Levy e Lemeshow, 2009). Traduzindo, é o mesmo que dizer que a probabilidade de qualquer elemento ser seleccionado é igual a $\frac{k}{N}$, o quociente entre a dimensão da amostra e a dimensão da população (as amostras em causa são retiradas da população sem reposição). Esta conclusão pode ser justificada matematicamente, tendo em conta o seguinte: o número de subconjuntos com k elementos que não contêm um determinado elemento é igual a $\binom{N-1}{k}$ ou seja o número de vezes em que os outros $N-1$ elementos se podem combinar em grupos de k elementos. A probabilidade de qualquer elemento não ser incluído é por isso igual $\dfrac{\binom{N-1}{k}}{\binom{N}{k}}$, que por sua vez é igual a $\dfrac{N-k}{N}$. A probabilidade de um determinado elemento ser seleccionado é portanto é igual a $1 - \dfrac{N-k}{N}$ ou $\dfrac{k}{N}$.

Este tipo de amostragem 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. De facto, na amostragem aleatória simples sem reposição todas as amostras de igual dimensão têm igual probabilidade de serem seleccionadas, uma equiprobabilidade que decorre da permutabilidade do esquema hipergeométrico.

Salienta-se, que a amostragem aleatória simples pode ser efectuada com reposição (caso em que cada elemento da população pode entrar mais do que uma vez na amostra), e nesse caso além de permutabilidade há independência entre as extracções sucessivas — mas perde-se a equiprobabilidade de todas as amostras de igual dimensão k, e a variância vem ampliada por um factor $\frac{N-1}{N-k}$.

Uma amostra aleatória simples pode ser recolhida mediante os seguintes passos:

1. Numerar os elementos da população de 1 a N.

2. Escolher aleatoriamente k números distintos de 1 a N, por recurso a uma tabela de números aleatórios.

3. Uma vez escolhidos os números, os elementos da população correspondentes constituem a amostra.

Amostragem Estratificada

Este método consiste em dividir a população em grupos relativamente homogéneos e mutuamente exclusivos, chamados estratos, e em seleccionar amostras aleatórias simples e 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.

Levy e Lemeshow (2009) definem amostragem estratificada como sendo um processo de amostragem em que a população é dividida em L estratos mútua e exaustivamente exclusivos, sendo retirada uma amostra aleatória de n_i elementos de cada estrato. É importante observar que a precisão de uma amostra não depende unicamente da dimensão da população, mas também da respectiva variabilidade. A variabilidade de um estrato é elevada quando os seus elementos têm características muito heterogéneas. Tal situação implica que um estrato com maior variância deverá levar à selecção de um maior número de unidades na amostra, quando comparado com um estrato com a mesma dimensão populacional mas menor variância (maior homogeneidade). A nível geográfico, por exemplo, os estratos mais urbanos apresentam comportamentos de consumo mais heterogéneos que os estratos com maior

índice de ruralidade, pelo que uma amostragem desproporcional permite obter dados mais rigorosos, através de uma sobre-amostragem nas regiões mais urbanas. Em resumo, quanto maior for o estrato, maior deve ser a amostra respectiva. Mas se a variabilidade dentro de um estrato for maior, maior deverá ser a respectiva sub-amostra.

A complexidade deste processo de amostragem exige os seguintes passos para passar à prática:

Definição dos Estratos

1. Para definição dos estratos utilizam-se vários métodos, entre eles, os seguintes: estudos piloto, informação de estudos anteriores, opiniões de conhecedores da população ou até mesmo recorrendo à intuição. A utilização de uma opinião subjectiva para identificar os estratos não vai por si só excluir esta amostra do grupo das amostras aleatórias (Hansen *et al.*, 1962), bem como não é necessariamente um factor que possa introduzir enviesamento nos resultados. A redução da variância dos estimadores (indicador da precisão dos resultados) pode ser expressa à custa de uma opinião fundamentada. Salienta-se que as variáveis geográficas, demográficas, económicas ou outras podem ser relevantes para o estudo do parâmetro em causa. Um aspecto a ter em conta é o número de estratos, pois quantos mais estratos se definirem, maior homogeneidade tendo em conta a relação do custo/orçamento disponível.

2. A selecção dos elementos dentro de cada estrato. Nesta fase a opção por uma estratificação proporcional ou não proporcional é determinista na escolha de quantos elementos de cada estrato se devem incluir na amostra. A amostra estratificada proporcional é aquela em que a proporção de elementos da amostra que possui determinadas características é idêntica à proporção de elementos da população que possui essas mesmas características (Bouquerel, 1974). Numa amostra estratificada proporcional em que a população tem tamanho N, em que se identificam k estratos com dimensões respectivas $N_1, N_2, \dots, N_k$, exprimindo-se matematicamente da seguinte forma:

$$N = N_1 + N_2 + N_3 + \cdots + N_k.$$

A média e a variância da população podem escrever-se em função das médias $\overline{Y}_i$ e das variâncias σ_i^2 de cada estrato, respectivamente da seguinte forma:

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{k} (N_i \times \overline{Y}_i) = \sum_{i=_1}^{k} (W_i \times \overline{Y}_i)$$

considerando que $W_i = \frac{N_i}{N}$, (se amostra estratificada proporcional) estabelecendo que W_i é o peso do estrato i na amostra, com $i = 1, \dots, k$. Note-se

que no caso em que a amostra estratificada não for proporcional, por exemplo por a heterogeneidade variar de estrato para estrato, as ponderações W_i são naturalmente afectadas pelo desvio padrão em cada um dos estratos.

Relativamente à variância define-se:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{N_i}(y_{ij}-\overline{y})^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{N_i}(y_{ij}-\overline{y}_i+\overline{y}_i-\overline{y})^2 =$$

$$= \sum_{i=1}^{k}\frac{N_i-1}{N-1}\times\sigma_i^2 + \sum_{i=1}^{k}\frac{N_i}{N-1}(\overline{Y}_i-\overline{Y})^2$$

3. A conjugação dos elementos seleccionados que constituem a amostra torna-se eficaz quando na população existem valores extremos para a característica em estudo, imputando a possibilidade de agregá-los num estrato separado. Administrativamente é conveniente a estratificação da amostra, pois a recolha de informação para uma amostra estratificada é análoga ao da amostra aleatória simples. Salienta-se, num caso destes, pouco há a perder em adoptar a amostra estratificada já que os respectivos desvios padrões resultantes raramente excederão os da amostra aleatória simples.

Amostragem por Clusters

Tal como na amostragem estratificada, na amostragem por clusters, a população é dividida em grupos, ou clusters. O termo cluster define um grupo de unidades elementares da população (Levy e Lemeshow, 2009). Este tipo de amostragem torna-se particularmente útil quando a população se encontra dividida num número reduzido de grupos, caracterizados por terem uma dispersão idêntica à população total, isto é, os grupos deverão, tanto quanto possível, ser "microcosmos "da população em estudo. Em primeiro lugar, 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 só exige apenas que se disponha de uma listagem dos grupos de indivíduos ou elementos da população. Um exemplo típico deste processo de amostragem remete-se a uma sondagem de opinião dos alunos de uma escola (população), da qual apenas se dispõe de uma listagem das turmas (grupos de alunos). Assim sendo, a amostra por clusters obtém-se seleccionando uma amostra aleatória de turmas e inquirindo, dentro de cada turma escolhida, todos os alunos.

A preferência pela amostra por clusters em muitos casos deve-se ao seu custo, geralmente inferior ao de outros métodos de amostragem, pelo menos no sentido em que o preço por unidade amostral observada é menor.

Amostragem Multi-Etapas

Este tipo de processo de amostragem é uma extensão da amostragem por clusters. Consideramos que a população está divida em vários grupos, e que posteriormente são seleccionados aleatoriamente alguns destes grupos. Este processo de selecção pode repetir-se por duas ou mais etapas se os grupos estiverem divididos em sub-grupos. A amostragem por muti-etapas segundo Scheaffer *et al.* (1990) é obtida mediante uma primeira selecção aleatória de clusters e depois, conforme os casos, seleccionar uma amostra de elementos ou então continuar com a selecção de clusters até chegar às unidades elementares.

Há vários processos que ilustram a aplicação deste tipo de amostragem, tais como :

1. Se pretendermos aferir a opinião dos alunos de qualquer instituição de ensino, dever-se-á começar por seleccionar aleatoriamente algumas das direcções escolares. Posteriormente em cada uma delas, seleccionar aleatoriamente algumas escolas, de cada uma das escolhidas seleccionar aleatoriamente algumas turmas e, por fim, em cada uma das turmas escolhidas seleccionar aleatoriamente alguns alunos. Neste exemplo há quatro etapas. Para a designação destas etapas é usual utilizar-se:
   (a) Unidade amostral primária — direcções escolares
   (b) Unidade amostral secundária — instituições de ensino (escolas)
   (c) Unidade amostral terciária — turmas
   (d) Unidade amostral quaternária — alunos (unidade estatística).
2. Pretende-se estimar a proporção de famílias que têm carros com uma determinada característica comum. O primeiro passo é considerar uma cidade previamente escolhida, o passo seguinte, será considerar a divisão da cidade em bairros, de entre os quais se seleccionaria uma amostra (unidade amostral primária). Dentro dos bairros seleccionados escolher-se-iam quarteirões (unidade amostral secundária), nos quarteirões escolhidos seleccionar-se-iam prédios (unidade terciária), e finalmente em cada prédio escolher-se-iam habitações (unidade quaternária). Salienta-se que neste processo de amostragem, os possíveis erros podem multiplicar-se, dado que ao longo deste processo se vão utilizando várias sub-amostras com a possibilidade de erros de amostragem em cada uma delas.

Amostragem Multi–Fases

Este processo de amostragem distingue-se do processo de amostragem multi-etapas, com o qual não deve ser confundido. Em cada fase de amostragem, consideram-se

sempre os elementos da população, mas as unidades amostrais variam de uma etapa para outra. Com a técnica multi-fásica, em cada fase de amostragem está sempre em causa o mesmo tipo de unidade amostral, obtendo-se de algumas unidades mais informação do que outras (Chisnall, 1986). No exemplo anteriormente exposto, respectivamente às escolas, as unidades amostrais eram, sucessivamente, as direcções escolares, as escolas, as turmas e os alunos, enquanto na amostragem multi-fásica se define sempre a mesma unidade amostral em todas as fases de extracção da amostra. A primeira fase assenta na recolha de dados sobre determinadas características, como por exemplo, o seu comportamento, frequência quanto ao consumo de um determinado produto, a disponibilidade para responder novamente a um inquérito. A informação do inquérito pode ser usada para a definição de uma listagem dos possíveis respondentes à segunda fase deste processo de amostragem. De seguida é então retirada da listagem uma segunda amostra que responderá a um questionário com um nível de profundidade mais elevado. Consequentemente, nem todos os inquiridos respondem a todas as questões, permitindo simultaneamente uma redução de custos e uma nova reutilização da amostra.

Este procedimento é particularmente eficaz na recolha de informação "sensível", em que o mesmo inquirido pouco a pouco vai enfraquecendo as barreiras na revelação de comportamentos íntimos, por exemplo.

2.3 *aassr* — Conceitos e Definições

Se considerarmos que a população tem dimensão N e quisermos uma amostra aleatória simples (sem reposição) de dimensão n, esta amostra é seleccionada aleatoriamente das $\binom{N}{n}$ amostras distintas possíveis, em cada uma das quais nenhum elemento da população é incluído mais do que uma vez. Isto é, cada uma das $\binom{N}{n}$ amostras possíveis tem a mesma probabilidade

$$\frac{n!(N-n)!}{N!}$$

Focando nas principais definições básicas da amostragem aleatória, o nosso interesse centra-se nos valores tomados por uma variável aleatória y que denota os elementos que são seleccionados para uma amostra

$$\boldsymbol{S} = (y_1, y_2, y_3, \ldots, y_n)\,,$$

onde y_i se refere ao i-ésimo elemento escolhido. A probabilidade de obter esta sucessão ordenada é : $\frac{1}{N} \times \frac{1}{N-1} \times \cdots \times \frac{1}{N-n+1} = \frac{(N-n)!}{N!}$. As características da população mais usualmente estudadas são:

1. O total da população $\tau = \sum_{k=1}^{N} y_k$

2. A média da população , $\mu = \sum_{k=1}^{N} \frac{y_k}{N} = \frac{\tau}{N}$

3. A proporção, P, representa a fracção de elementos da população que pertencem a uma determinada categoria de classificação da variável Y. Especificando um exemplo, uma fracção da amostra constituída por 2898 alunos, com idades compreendida dos 12 e 18 anos de escolas da Área Metropolitana do Porto e de Lisboa, revela comportamentos desviantes. A conclusão de um estudo realizado pela Escola de Criminologia e apresentada no âmbito do 72º Curso Internacional de Criminologia e posteriormente publicado pelo jornal "Metro, Sexta-feira, 26 de Novembro de 2010" lê-se o seguinte: "metade dos alunos das áreas metropolitanas, entre os 12 e 18 anos, já ingeriu bebidas fortes, um em cada dez usou drogas e há uma ligação 'forte' em 91.4% dos casos, entre estes consumos e comportamentos desviantes". O objectivo de estudo por amostragem é estimar uma ou mais dessas categorias a partir da informação contida na amostra de n ($\leq N$) membros da população, sendo $P = \frac{n}{N}$ a fracção que deve permitir construir o intervalo de confiança de prevalência destes comportamentos desviantes na população.

Média e Variância

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

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

Numa amostra aleatória simples o valor esperado de y_i, a i-ésima observação, define-se aplicando-se o seguinte modelo matemático:

$$E[y_i] = \sum_{j=0}^{N} Y_j \, \mathbb{P}(y_i = Y_j) = \frac{1}{N} \sum_{j=1}^{N} Y_j = \overline{Y},$$

pois $\mathbb{P}(y_i = Y_j) = \frac{1}{N}$ corresponde ao facto de que o número de amostras em que $y_i = Y_j$, e cada uma tem probabilidade de $\frac{(N-n)!}{N!}$.

Analogamente verifica-se que

$$E[y_i^2] = \sum_{j=1}^{N} Y_j^2 \, \mathbb{P}\left[y_i = Y_j\right] = \frac{1}{N} \sum_{j=1}^{N} Y_J^2,$$

E consequentemente

$$Var(y_i) = E[(y_i - \overline{Y})^2] = E[y_i^2] - \overline{Y}^2 = \frac{(N-1)\sigma^2}{N}.$$

Probabilidade Conjunta e Covariância

No que se refere à probabilidade conjunta, matematicamente tem-se o seguinte modelo de probabilidade:

$$P(y_i = Y_r, y_j = Y_s) = (N-1)\,(N-2)\,(N-3)\,\ldots\,(N-n+1)\,\frac{(N-n)!}{N!} = \frac{1}{N(N-1)},$$

$r \neq s$.

Consequentemente, de

$$\mathbb{E}\left(y_k y_j\right) = \frac{2}{N(N-1)} \sum_{r<s} Y_r Y_s = \frac{1}{N(N-1)} \sum_{r\neq s} Y_r Y_s =$$

$$= \frac{1}{N(N-1)} \left[\sum_{r=1}^{N} \sum_{s=1}^{N} Y_r\, Y_s - \sum_{i=1}^{N} Y_i^2 \right] =$$

$$= \frac{1}{N(N-1)} \left[\left(\sum_{i=1}^{N} Y_i \right)^2 - \sum_{i=1}^{N} Y_i^2 \right]$$

segue-se que

$$\mathrm{Cov}\left(y_k y_j\right) = \mathbb{E}\left(y_k y_j\right) - \mu^2 =$$

$$= \frac{1}{N(N-1)} \left[\left(\sum_{k=1}^{N} Y_k \right)^2 - \sum_{j=1}^{N} Y_j^2 - N(N-1)\mu^2 \right] =$$

$$= \frac{1}{N(N-1)} \left[N^2\mu^2 - \sum_{j=1}^{N} Y_j^2 - N(N-1)\mu^2 \right] =$$

$$= -\frac{1}{N(N-1)} \left[\sum_{j=1}^{N} Y_j^2 - \mu^2 \right] = -\frac{1}{N}\,\sigma^2.$$

A correlação é naturalmente negativa, uma vez que a extracção sem reposição de um elemento de qualquer uma das classes em que a população está particionada reduz a probabilidade de na extracção seguinte se retirar um elemento dessa classe. Por outro lado, a permutabilidade, característica de extracções sem reposição, é uma dependência fraca, e observa-se que a covariância — e consequentemente a correlação -$\frac{1}{N-1}$ —, sendo $O(\frac{1}{N})$, é obviamente pequena.

Estimação — Generalidades

Um estimador $\Theta = \Theta(X_1, \ldots, X_n)$ é uma estatística (uma variável aleatória função da amostra aleatória, apenas, que consequentemente não depende de parâmetros desconhecidos) cujas realizações fornecem avaliações (estimativas) do parâmetro desconhecido.

Representamos uma certa população ou universo por X. O comportamento de X é conhecido quando se conhece a família parametrizada das distribuições adequada, e o valor dos parâmetros caracterizadores dessa distribuição. Suponha-se que θ é um parâmetro desconhecido, e $f(x;\theta)$ a densidade — massa de probabilidade no caso discreto, função densidade de probabilidade no caso absolutamente contínuo, para nos centrarmos nos casos elementares que são, afinal, os que na generalidade das situações ajustamos à realidade — da variável X, expressa em função do parâmetro desconhecido.

A estimação de θ pode ser feita usando metodologias diversas (método dos momentos, método da verosimilhança máxima, por exemplo), encontrando-se em Casella e Berger (2002, pp. 271–299) uma excelente exposição dos grandes princípios orientadores — suficiência, verosimilhança, equivariância — que devem enformar a redução dos dados e a inferência estatística, por outras palavras a alquimia da transformação de informação em conhecimento. Qualquer que seja a escolha de princípios orientadores e metodologias de redução dos dados, podemos estar interessados numa mera estimativa pontual, precisa mas de probabilidade nula, ou num intervalo de confiança com um grau de probabilidade elevado, sendo o preço a pagar uma imprecisão, que na generalidade das situações interessantes pode ser controlada através de uma amostragem adequada.

- Estimação pontual: produção de um valor, que se pretende que seja o melhor[(1)] para um determinado parâmetro da população, com base na informação amostral. Ou seja, o objectivo da estimação pontual é produzir uma estimativa de θ, que pertença ao conjunto de valores admissíveis (espaço dos parâmetros) que o parâmetro pode assumir, de acordo com a distribuição de X. Por exemplo, se X segue uma Binomial, $X \frown B(n,p)$, os parâmetros da população desconhecidos são o n e o p, exigindo-se as condições seguintes: $0 \leq p \leq 1$ e $n \in \mathbb{N}^+$.

- Estimação intervalar: construção de um intervalo que, com certo grau de confiança previamente definido, contenha o verdadeiro valor do parâmetro da população. A construção de intervalos de confiança parte de uma variável fulcral que permite construir um estimador intervalar com a probabilidade de cobertura pretendida, e tendo em vista a dualidade de estimação intervalar e

(1) Veja-se por exemplo em Pestana e Velosa (2010, pp. 511–512) uma discussão elementar de suficiência, centralidade, eficiência, consistência e robustez de um estimador pontual.

testes de hipóteses bilaterais, em algumas situações recorre-se directamente a um intervalo obtido por reinterpretação da regia de manutenção da hipótese nula do teste bilateral adequado.

Estimador da Média $\overline{Y}$ e sua Variância

Teorema 2.3.1.

Na amostragem aleatória simples sem reposição

1. *O estimador* $\overline{y}$ *é centrado;*
2. *o erro padrão de* $\overline{y}$ *é* $\sqrt{\frac{s^2}{n}(1-f)}$*, onde* $f = \frac{n}{N}$ *é a fracção amostral.*

Demonstração:

Quanto a 1., basta recordar o já anteriormente apontado:

$$\mathbb{E}(\overline{y}) = \mathbb{E}(y_k) = \sum_{k=1}^{N} Y_k \, \mathbb{P}[y_k = Y_k] = \sum_{k=1}^{N} Y_k \, \frac{1}{N} = \mu \,,$$

porque há $\frac{(N-1)!}{(N-n)!}$ amostras em que $y_k = Y_k$, cada uma das quais com probabilidade $\frac{1}{\binom{N}{n}}$, pelo que $\mathbb{P}[y_k] = Y_k = \frac{(N-1)!}{(N-n)!}\,\frac{1}{\binom{N}{n}} = \frac{1}{N}$.

O cálculo da variância do estimador $\overline{y}$ é agora imediato, recordando que a variância da soma é a soma das covariâncias (a variância de cada parcela é o caso especial da covariância dessa parcela consigo mesma):

$$\mathrm{Var}(\overline{y}) = \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{k=1}^{n} y_k\right) = \frac{1}{n^2}\,\sum_{k=1}^{n}\mathrm{Var}(y_k) + \frac{1}{n^2}\,\sum_{r\neq s}\mathrm{Cov}(y_r y_s) =$$

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

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

onde $f = \frac{n}{N}$ é a fracção amostral. □

A expressão $\frac{N-n}{N} = 1 - f$ é o *factor de correcção para populações finitas.*

Embora ocasionalmente usemos estimadores enviesados, é natural preferir estimadores centrados. E, entre estes, desejar eficiência máxima, por outras palavras que os estimadores estejam associados a variáveis fulcrais que permitam, para um dado coeficiente de confiança, obter intervalos tão curtos quanto possível. Vamos mostrar que o estimador $\overline{y}$ é, a este respeito, óptimo:

Teorema 2.3.2.

O estimador $\overline{y}$ é o estimador linear centrado de variância mínima de μ.

Demonstração:

Considere-se a forma geral dos estimadores lineares de μ,

$$T = \sum_{k=1}^{n} \alpha_k y_k, \qquad \text{com} \qquad \sum_{k=1}^{n} \alpha_k = 1$$

a fim de T ser centrado. A variância de T é

$$\begin{aligned}
\mathrm{Var}(T) &= \frac{N-1}{N}\,\sigma^2 \sum_{k=1}^{n} \alpha_k^2 - \frac{\sigma^2}{N} \sum_{r \neq s} \alpha_r \alpha_s = \\
&= \frac{N-1}{N}\,\sigma^2 \sum_{k=1}^{n} \alpha_k^2 - \frac{\sigma^2}{N} \left[\left(\sum_{k=1}^{n} \alpha_k \right)^2 - \sum_{k=1}^{n} \alpha_k^2 \right] = \\
&= \left(\sum_{k=1}^{n} \alpha_k^2 - \frac{1}{N} \right) \sigma^2,
\end{aligned}$$

que é tanto menor quanto menor for $\sum_{k=1}^{n} \alpha_k^2$ sujeito à condição $\sum_{k=1}^{n} \alpha_k = 1$.

Queremos então minimizar

$$\sum_{k=1}^{n-1} \alpha_k^2 + \left((1 - \sum_{k=1}^{n} \alpha_k \right)^2 .$$

Por outras palavras, exigimos

$$\frac{\partial}{\partial \alpha_j} \sum_{k=1}^{n-1} \alpha_k^2 = \frac{\partial}{\partial \alpha_j} \left[- \left((1 - \sum_{k=1}^{n} \alpha_k \right)^2 \right] \iff \alpha_k = 1 - \sum_{k=1}^{n-1} \alpha_k = \alpha_n,\ k = 1, \ldots, n-1,$$

e como $\sum_{k=1}^{n} \alpha_k = 1$, segue-se que para T ser o estimador linear centrado de variância mínima os coeficientes $\alpha_k = \frac{1}{n}$, $k = 1, 2, \ldots, n$, isto é $T = \overline{y}$. □

Como $\text{Var}(\overline{y}) = \frac{\sigma^2}{n}(1-f) \xrightarrow[n\to\infty]{} 0$, $\overline{y}$ é um estimado consistente de μ.

Estimação da Variância σ^2

A $var(\overline{y})$ é aplicada de três formas distintas:

1. Para investigar a precisão do estimador $\overline{y}$ de $\overline{Y}$;
2. Para comparar $\overline{y}$ com outros estimadores de $\overline{Y}$, nomeadamente no que respeita à eficiência;
3. Para determinar a dimensão da amostra necessária para obtenção da precisão do estimador $\overline{y}$ que se pretende para o estudo em causa.

Normalmente, não se conhece o valor σ^2, como tal é necessário estimá-lo a partir da amostra. A variância empírica (por vezes chamada variância corrigida) definida como habitualmente é um estimador centrado da variância populacional σ^2, pois

$$E(\tilde{S}^{\,2}) = E\left(\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-1}\right) = \frac{1}{n-1}\,E\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right) =$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} E(y_i)^2 - \frac{n}{n-1}\,E(\overline{y}^2) = \frac{1}{n-1}\left[n\,\overline{Y} + n\,\sigma^2\right] - \frac{n}{n-1}\left[\overline{Y} + \frac{\sigma^2}{n}\right] = \sigma^2$$

e é o estimador da variância geralmente usado.

Estimação Intervalar — Generalidades

Os intervalos de confiança são hoje uma prática comum na comunicação social para a divulgação de qualquer estudo, seja ele de que natureza for. Em qualquer sondagem efectuada, indica para além das estimativas pontuais, uma ficha técnica em que os intervalos de confiança são indicados.

Dada uma grande variedade de problemas de inferência o interesse não se restringe a estimar um parâmetro, mas sim estabelecer um limite inferior ou superior,

ou ambos, para o parâmetro, ou seja, construir uma família de intervalos de confiança de tal forma que uma elevada proporção destes possa conter o parâmetro. O ideal é construir intervalos de confiança com elevada probabilidade de conter o parâmetro, mas que tenham simultaneamente amplitudes muito pequenas.

Metodologia Aplicável

A metologia para a construção de um intervalo de confiança para um dado parâmetro θ é a seguinte:

1. Definição da população, da sua distribuição com uma parametrização adequada e do parâmetro a estimar;
2. Escolha da variável fulcral, ou seja, da função da amostra aleatória e do parâmetro que se pretende estimar que vamos utilizar.
3. Determinação da distribuição da variável fulcral;
4. Escolha de nível de significância α, ou do nível de confiança $(1-\alpha)\times 100\%$;
5. Determinação dos limites do intervalo de confiança, a partir dos valores da amostra.

Intervalos de Confiança para $\overline{Y}$

A grande limitação dos métodos de estimação é a de não fornecerem qualquer informação relativa ao rigor das estimativas efectuadas. Esta dificuldade é ultrapassada recorrendo aos intervalos de confiança. Só é possível construir um intervalo de confiança para um parâmetro se for conhecida a distribuição do estimador intervalar utilizado. Recorre-se ao Teorema Limite Central de Erdös–Rényi (para a situação de permutabilidade, uma vez que centrámos a nossa atenção na amostragem sem reposição). A aproximação pela distribuição normal standard é em geral admitida, sem grandes pruridos no que se refere à velocidade de convergência[(2)]. Então, como a média amostral, $\overline{y}$, de uma amostra aleatória simples tem aproximadamente distribuição normal,

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

[(2)] Que é lenta, o resultado geral é que é $O(\frac{1}{\sqrt{n}})$, mas sabe-se que quando a população parente tem terceiro cumulante pequeno — e portanto não é muito assimétrica — a velocidade de convergência é muito melhor, $O(\frac{1}{n})$.

Este resultado é consensual e bastante razovável, mesmo quando existe assimetria na população. Usualmente uma regra empírica para a utilização de aproximação à normal é exigir que a dimensão da amostra satisfaça

$$n > 25G_1^2 \qquad \text{onde} \qquad G_1 = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(Y_i - \overline{Y})^3.$$

Salienta-se que para populações finitas G_1 é análogo ao coeficiente de assimetria de Fisher. Sublinha-se que o factor de amostragem, $f = \frac{n}{N}$ não deve ser elevado. Para um $n > 40$ o intervalo de confiança a $(1-\alpha)\times 100\%$ para $\overline{Y}$ é apresentado da seguinte forma:

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

Na prática, o valor de σ^2 não é conhecido e tem que se utilizar a estimativa s^2, pelo que se obtém um intervalo de confiança aproximado.

No caso em que o valor de $n \leq 40$ é aconselhável utilizar a distribuição t de Student para o intervalo de confiança a $(1-\alpha)\times 100\%$ que é dado por

$$\left]\overline{y} - t_{n-1,1-\frac{\alpha}{2}}\,S\sqrt{\frac{1-f}{n}}; \overline{y} + t_{n-1,1-\frac{\alpha}{2}}\,S\sqrt{\frac{1-f}{n}}\right[$$

estabelecendo que $t_{n-1,1-\frac{\alpha}{2}}$ é o quantil de probabilidade $1-\frac{\alpha}{2}$ da distribuição de Student com $n-1$ graus de liberdade.

Anote-se porém que a sitação geral é determinar n por forma a ter um intervalo de confinaça apertado ($< B$), como adiante discutimos, e por isso não faz muito sentido esta última observação. De facto, na prática, usam-se sempre os quantis relevantes da aproximação pela gaussiana.

Dimensão da Amostra

Ao projectar uma sondagem, uma questão que se pretende resolver desde o início é a decisão de quantos indivíduos a amostra deve conter. É uma decisão que não é fácil de tomar de ânimo leve, pois na sua essência há que contrabalançar dois efeitos: a precisão, que à partida aumenta com a dimensão da amostra, e o custo, que também aumenta na mesma direcção. Uma vez clarificado o que se pretende com o estudo, este prossegue com as decisões a tomar referentes ao processo e enquadramento amostral:

1. Características da população, nomeadamente a variância da característica em estudo e o número de elementos que corresponde à sua dimensão $\boldsymbol{N}$;
2. Distribuição amostral do estimador a utilizar;
3. Precisão e confiança requeridas para os resultados, especificando a diferença máxima entre a estimativa e o parâmetro ou nível de confiança;
4. Custo elevado devido à recolha em demasia de inquéritos;
5. Ilação para os erros de não amostragem;

> A validade de uma sondagem é função do seu erro total (Assael e Keon, 1982). É uma questão diíicil para o investigador na realização do estudo por sondagem, pois a grandeza da amostra a seleccionar recai na redução ou não do erro amostral, ou se deve concentrar nos recursos e esforços numa amostra de dimensão mais reduzida, mantendo um estudo com qualidade. O ideal será estabelecermos a precisão desejada, ou o gasto máximo que podemos realizar, e escolher a dimensão da amostra em função destas restricões;

Técnicas Estatísticas para Determinação da Dimensão

O desconhecimento de alguns parâmetros da população afecta a precisão dos estimadores.

Assumindo, por simplicidade de exposição, que o objectivo é estimar apenas uma característica, a média da população, $\overline{Y}$, utilizando a média amostral $\overline{y}$, e impondo que a probabilidade da diferença absoluta entre $\overline{Y}$ e $\overline{y}$ ser superior a um dado valor não exceda um certo nível de probabilidade, interessa determinar o valor mínimo de n que assegure $P(|\overline{Y} - \overline{y}| > d) \leq \alpha$.

Reduzindo,

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

para utilizar a aproximação à distribuição normal[(3)], obtém-se:

$$\frac{d}{\sigma\sqrt{(1-f)/n}} \geq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \Longleftrightarrow$$

[(3)] Uma vez que estamos no contexto de extracções sem reposição, a aproximação pela gaussiana não pode ser justificada pelo Teorema Limite Central clássico, em que se assume que as parcelas são independentes. Existe no entanto uma extensão do Teorema Limite Central para o caso de parcelas permutáveis, de Erdös e Rényi, veja-se Pestana e Velosa (2010, pp. 1036–1046), que plenamente justifica a aproximação em causa. Não discutimos a problemática da velocidade de convergência desse resultado, mas sabe-se que uma estimativa por defeito é ser um $O\left(\frac{1}{\sqrt{n}}\right)$, e consequentemente parece uma aproximação aceitável nas condições práticas usuais.

$$\iff n \geq \left(1 - \frac{n}{N}\right) \times \left(\frac{\sigma \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)}{d}\right)^2$$

e portanto

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

Como

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

a desigualdade acima pode ser reescrita da seguinte forma:

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

e assim uma aproximação aceitável para o valor pretendido de n é :

$$n_0 = \frac{\sigma^2}{V} \qquad (0.1)$$

Porém o valor encontrado para n deve ser avaliado tendo em conta os recursos disponíveis para a realiazação da amostragem, pois a expressão encontrada (0.1) avalia por excesso a dimensão da amostra. Isto exige uma estimativa de custo, tempo, pessoal e material necessário para cumprir a dimensão estabelecida, para tal é necessário reduzir a aproximação, considerando

$$n = \frac{n_0}{1 + \frac{n_0}{N}},$$

pressupondo que σ^2 é conhecido. Reportanto para a prática tal não acontece, assim sendo, é necessário estimar a dimensão da amostra requerida tomando o σ^2 desconhecido. Para estimar existem várias formas de o fazer, destacando-se entre elas quatro formas:

1. Recorrendo a um estudo piloto;
2. Recorrendo a inquéritos efectuados;
3. Usando resultados de uma amostra preliminar;
4. Recorrendo a considerações práticas acerca da estrutura da população.

Estimação de uma Proporção

Uma sondagem pode ser realizada para determinar a proporção de elementos na população que possuem, ou não, uma determinada característica. A estimação de uma proporção toma por base uma população de Bernoulli, na qual uma observação ou pertence ou não pertence à classe cuja proporção se pretende estimar. O estimador $\boldsymbol{P}$ dessa proporção de elementos que caem na categoria que interessa estudar é definido por $\widehat{P}$, e é afinal uma média amostral $\overline{X}$ — pelo que a teoria facilmente se reporta à anteriormente exposta — se as observações foram definidas da seguinte forma:

$$X_j = \begin{cases} 0, & \text{se o } j\text{-ésimo elemento da amostra não possui o atributo especificado} \\ 1, & \text{se o possui} \end{cases}$$

Assim sendo, o total de elementos na amostra que verifica uma determinada característica específica é dado por

$$X_T = \sum_{j=1}^{N} X_j.$$

Portanto, a proporção amostral $\widehat{P}$ é a fracção dos elementos da amostra que possuem o atributo de interesse para o estudo, ou seja,

$$\widehat{P} = \sum_{j=1}^{N} \frac{X_j}{n} = \overline{X}.$$

Salienta-se que existe uma particularidade para o valores de $\widehat{P}$, tendo em consideração que $\widehat{P}$ é média de valores 0 e 1 da amostra. Consequentemente implica a existência de uma relação entre $\overline{X}$ e σ_X^2. Pela condição anteriormente matemáticamente verifica-se que:

$$\sigma^2 = \frac{1}{N-1}\sum_{j=1}^{N}(X_j - \overline{X})^2 = \frac{1}{N-1}\sum_{j=1}^{N}(X_j - P)^2 = \frac{1}{N-1}\sum_{j=1}^{N} X_j^2 - \frac{N}{N-1}\cdot P^2 =$$

$$= \frac{1}{N-1}\sum_{j=1}^{N} X_j - \frac{N}{N-1}\cdot P^2 = \frac{N\cdot P(1-P)}{N-1}$$

A resolução desta igualdade implica o conhecimento de

$$V(\widehat{P}) = (1-f)\cdot\frac{\sigma_X^2}{n} = \frac{N-n}{n(N-1)}\cdot P\cdot(1-P),$$

ou seja, $V(\overline{X})$, também se verifca que $E(P) = P$, isto é, P é centrado.

Quando o P é desconhecido o problema coloca-se em conhecer a variância da população, logicamente podemos estimar σ_X^2 pelo estimador centrado, demonstrando que :

$$S^2(P) = Var(\widehat{P}) = \frac{N-n}{n} \cdot \frac{P(1-P)}{N-1} = \frac{N-n}{N} \cdot \frac{N}{n(N-1)} \cdot P(1-P) =$$

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

do que resulta

$$S^2(P) = (1-f) \cdot \frac{P(1-P)}{n-1},$$

e consequentemente é um estimador centrado de $Var(\widehat{P})$.

Intervalos de Confiança para uma Proporção

Havendo R elementos da população com o atributo, então a probabilidade de na amostra se observarem r elementos com atributo, tem como distribuição exacta a distribuição hipergeométrica

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

ou seja, o número de elementos da amostra, de dimensão n, com o atributo em causa tem distribuição hipergeométrica com parâmetros $N, n, \frac{R}{N}$. Evidentemente pode-se construir intervalos de confiança para P, dado que se cnnhece a sua distribuição.

Na prática a construção do intervalo de confiança para P, envolve cálculos trabalhosos, que na prática não é geralmente utilizada. A aproximação da distribuição hipergeométrica à binomial, é naturalmente aceitável se $f = \frac{n}{N} \simeq 0$ e $(N-R)$, pode-se ignorar tendo em conta que estamos perante uma amostra extraída sem reposição. Considerando r o número de elementos da amostra com atributo e $r < n$ tem-se uma distribuição aproximadamente $B(n,p)$, ou seja,

$$P_r = P(Y=r) = \frac{\binom{N}{r} \cdot \binom{N-R}{n-r}}{\binom{N}{n}} = \frac{\binom{N_p}{r} \cdot \binom{N_q}{n-r}}{\binom{N}{n}} =$$

$$= \binom{N}{r} \frac{Np(Np-1)\ldots[Np-(r-1)]Nq(Nq-1)\ldots[Nq-(n-r-1)]}{N(N-1)\ldots(N-(n-1))} \approx$$

$$\approx \binom{n}{r} p^r (1-p)^{n-r} = P[X=r],$$

onde

$$X \frown Binomial(n, p).$$

Contudo, a aplicação desta distribuição para a elaboração de intervalos de confiança envolve métodos de cálculo pesados. A alternativa conveniente é a aproximação à distribuição normal, nas seguintes condições:

1. n não muito grande relativamente a R ou $N - R$;
2. $\min(np, n(1-p)) > 30$.

formulando que

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

obtém-se o seguinte:

$$\frac{p-P}{\sqrt{(1-f) \cdot \frac{p(1-p)}{n}}} \sim N(0,1)$$

assim sendo, a construção do intervalo de confiança a $(1-\alpha) \times 100\%$ para P é elaborado da seguinte forma:

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

sendo que

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

Quando n for suficientemente grande, a simplificação fica mais reduzida, fazendo a substituição de $Var(P)$ pelo seu estimador centrado $S^2(P)$, obtendo-se o intervalo com o mesmo nível de confiança da seguinte forma:

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

o mesmo é

$$P \pm z_{1-\frac{\alpha}{2}}\,\sqrt{(1-f)\frac{p(1-p)}{n-1}}$$

onde $z_{1-\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$ é o quantil da distribuição normal padrão.

Determinação da Dimensão que Deve Ter uma Amostra Aleatória

A questão que em geral nos interessa é, naturalmente, a perspectiva inversa, isto é, se quisermos estimar P com um erro máximo d, mais precisamente,

$$P\left(|p-P|>d\right)\leq\alpha,$$

considerando a aproximação à normal, resulta no seguinte:

$$z_{1-\frac{\alpha}{2}}\sqrt{(1-f)\frac{\sigma^2}{n}}\ <\ d\iff\ \frac{1}{n}\left(1-\frac{n}{N}\right)\ <\ \left(\frac{d}{\sigma z_{1-\frac{\alpha}{2}}}\right)^2\iff$$

$$\iff\frac{1}{n}<\frac{1}{N+\left(\frac{d}{\sigma\ z_{1-\frac{\alpha}{2}}}\right)^2}\iff n>N+\left(\frac{d}{\sigma\ z_{1-\frac{\alpha}{2}}}\right)^2=\ N\ \frac{\sigma^2}{\sigma^2+N\cdot D^2}$$

onde $D=\frac{d}{z_{1-\frac{\alpha}{2}}}$, isto é equivalente a

$$n>\left(\frac{\sigma}{D}\right)^2\cdot\frac{1}{\frac{1}{N}\cdot\left(\frac{\sigma}{D}\right)^2+1}\approx\left(\frac{\sigma}{D}\right)^2=n_0.$$

Como primeira aproximação considera-se

$$n_0=\frac{P(1-P)}{\left(\frac{d}{z_{1-\frac{\alpha}{2}}}\right)^2}.$$

No caso da expressão obtida não ser negligível, esta sobreavaliação da expressão pode ser escrita

$$n\geq\frac{n_0}{1+\frac{n_0-1}{N}}.$$

Estimação do Total da População

Em muitas situações a estimação do total da população Y_T torna-se desnecessária, considerando que $Y_T=N\overline{Y}$. Através da relação entre Y_T e $\overline{Y}$ facilmente se deduzem as propriedades sobre estimação populacional total. Dado que estamos na presença de uma amostragem aleatória simples o estimador mais utilizado é dado por

$$y_T=N\overline{y},$$

sendo o mesmo centrado, basta notar que

$$E(\overline{y}) = E(y_j) = \sum_{j=1}^{N} Y_j \ P[y_j = Y_j] = \sum_{j=1}^{N} Y_j \frac{1}{N} = \mu,$$

sendo que a probabilidade

$$P[y_j] = Y_j = \frac{(N-1)!}{(N-n)!} \frac{1}{\binom{N}{n}} = \frac{1}{N},$$

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

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

Demonstra-se que y_T é um estimador linear centrado de variância mínima. Para um $n > 40$ com um valor de fracção amostral próximo de zero, pode-se usar a aproximação à distribuição normal

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

para construir intervalos com coeficiente de confiança aproximadamente $(1-\alpha) \times 100\%$

$$\left(y_T - z_{1-\frac{\alpha}{2}} \ \sigma \ N \ \sqrt{\frac{1-f}{n}}, y_T + z_{1-\frac{\alpha}{2}} \ \sigma \ N \ \sqrt{\frac{1-f}{n}}\right).$$

Para um $n \leq 40$, é conselhável utilizar o quantil $t_{n-1,1-\frac{\alpha}{2}}$ em vez do quantil $z_{1-\frac{\alpha}{2}}$ da distribuição normal reduzida.

Em relação à escolha da dimensão da amostra,tem-se em conta o seguinte

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

Utilizando a aproximação pela normal vem que

$$n \geq \ N \left[1 + \frac{1}{N}\left(\frac{d}{\sigma \ z_{1-\frac{\alpha}{2}}}\right)^2\right]^{-1},$$

sendo equivalentemente,

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

a expressão pode ser reescrita

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

se $\frac{n\sigma^2}{V} \simeq 0$ pode considerar-se

$$n_o = \frac{n^2\sigma^2}{V},$$

e no caso contrário deve-se utilizar

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

2.4 O Estimador de Horvitz–Thompson

Expusemos com algum detalhe a amostragem aleatória simples sem reposição. Os estimadores da média e do total de uma população, e de uma proporção (que é uma média especial, de zeros e uns) podem ser perspectivados como caso especial de um estimador genérico proposto por Horvitz and Thompson (1952), apropriado também para fazer o enquadramento geral de estimadores daqueles parâmetros amostrais em outras metodologias que descrevemos sucintamente. Expomos por isso, sem detalhe excessivo, a teoria de Horvitz e Thompson, que na sua essência é simplesmente usar coeficiente de ponderações adequados para expandir o que se observa na amostra para a população.

Denotamos $\mathcal{P}$ uma população de dimensão N, e $\mathcal{S}$ um plano amostral, isto é, uma função que a cada subconjunto $\boldsymbol{s} \subset \mathcal{P}$ atribui uma probabilidade $\mathbb{P}(\boldsymbol{s})$ de ser seleccionado como amostra.

Consideramos apenas planos amostrais *próprios*, aqueles em que a probabilidade $\pi_{_i}$ de selecção do i-ésimo elemento da população como elemento da amostra, $\forall i \in \{1,\ldots,N\}$, é positiva. Os recíprocos das probabilidades de inclusão, $w_{_i} = \frac{1}{\pi_{_i}}$, são os *pesos de amostragem*.

Para simplicidade de exposição, admita-se que nos interessa estimar a média populacional μ da variável Y. Denotamos os seus valores na população $Y_{_1}, Y_{_2}, \ldots, Y_{_N}$, e os seus valores na amostra $y_1, y_2, \ldots, y_n$. Note-se que a dimensão n da amostra pode ser um valor fixo ou variável e, neste caso, inclusivamente aleatório. A aleatoriedade da amostra decorre apenas das probabilidades de inclusão $\pi_{_i} = \mathbb{P}\left[Y_{_i} = y_{_k} \in \boldsymbol{s}\right]$.

Denotando $\mathcal{I}(i)$ o indicador de que o i-ésimo membro da população é seleccionado para a amostra, a probabilidade de inclusão é

$$\pi_{_i} = \mathbb{E}\left[\mathcal{I}(i)\right].$$

Por outro lado, a probabilidade de inclusão $\pi_{_i}$ é a probabilidade dos subconjuntos de $\mathcal{P}$ de que esse elemento é membro:

$$\pi_{_i} = \sum_{\left\{\boldsymbol{s}\subset\mathcal{P}:\, Y_{_i}\in\boldsymbol{s}\right\}} \mathbb{P}\left(\boldsymbol{s}\right).$$

Esta expressão geral é útil nas situações em que o plano amostral implica simetrias que tornam aquele cálculo fácil, por exemplo em *aassr* a probabilidade de inclusão de qualquer elemento de uma população de dimensão N numa amostra de dimensão n é $\pi_{_{i^\star}} = \pi = \frac{n}{N}$.

Retomando o problema de estimar a média populacional $\mu = \frac{1}{N}\sum_{i=1}^{N} Y_{_i}$ com base na amostra $(y_{_1}, \ldots, y_{_n})$. O estimador[(4)] intuitivo de μ é a média amostral $\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_{_i}$ que pode facilmente ser expresso em termos dos valores populacionais:

$$\tilde{\mu} = \frac{\sum_{i=1}^{N} \mathcal{I}(i)\, Y_{_{i^\star}}}{\sum_{i=1}^{N} \mathcal{I}(i)}$$

Admita-se, de momento, que a dimensão da amostra é um valor fixo n (isto é, que apenas subconjuntos de dimensão n têm probabilidade positiva de ser seleccionados). Neste pressuposto, e denotando $\pi_{_{ij}}$ a probabilidade de inclusão conjunta dos i-ésimo e j-ésimo elementos da população na amostra, com a notação simplificada $\pi_{_i} = \pi_{_{ii}}$,

$$\mathbb{E}\left(\tilde{\mu}\right) = \frac{1}{n}\sum_{i=1}^{N} \pi_{_i}\, Y_{_i}$$

e

$$n^2 \operatorname{var}(\tilde{\mu}) = \sum_{i=1}^{N} \pi_{_i}\left(1-\pi_{_i}\right) Y_{_i}^{^2} + \sum_{i=1}^{N}\sum_{j=1}^{N} (\pi_{_{ij}} - \pi_{_i}\pi_{_j})\, Y_{_i}\, Y_{_j} =$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} \pi_{_{ij}}\, Y_{_i}\, Y_{_\jmath} - \left(\sum_{\ell=1}^{N} \pi_{_\ell}\, Y_{_\ell}\right)^{^2}.$$

(4) Usamos o termo estimador e não estimativa porque consideramos que se trata de uma amostra aleatória, no sentido de haver probabilidades de inclusão.

O viés B do estimador $\tilde{\mu}$ é

$$B(\tilde{\mu}) = \frac{1}{n} \sum_{i=1}^{N} \left(\pi_{_i} - \frac{\mathbb{E}(n)}{N} \right) Y_{_i}$$

quando a variável indicatriz $\mathcal{I}$ da inclusão dos elementos da população na amostra e Y são correlacionadas. Note-se, consequentemente, que $\tilde{\mu}$ é centrado quando no plano amostral há equiprobabilidade de inclusão de qualquer elemento da população na amostra de dimensão fixa n. Observe-se, contudo, que o preço da amostragem é um argumento que, em geral, tem precedência nas decisões sobre o plano amostral, e que há planos amostrais que proporcionam uma boa eficiência (no sentido de erro quadrático médio moderado) do estimador $\tilde{\mu}$.

Quando o denominador é aleatório o cálculo do valor médio e da variância deste estimador pode ser extremamente complicado, obrigando em geral a recorrer ao método delta (Diamantino, 2008; Martins, 2009).

Como $\mathbb{E}\left[\mathcal{I}(i)\, w_{_i}\right] = 1$, da expressão $\mathbb{E}\left(\tilde{\mu}\right) = \frac{1}{n} \sum_{i=1}^{N} \pi_{_i}\, Y_{_i}$ conclui-se que se $\mathcal{I}$ e Y forem não correlacionadas o estimador de Horwitz-Thompson

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{I}(i)\, w_{_i}\, Y_{_i} \tag{2.1}$$

é centrado.

Supondo que $\boldsymbol{s}$, $Y_{_i} = y_{_k}$, isto é que $Y_{_i}$ é o k-ésimo elemento da população $\mathcal{P}$ seleccionado para a amostra, e denotando $\pi_{_i} = \mathbb{P}\left[\mathcal{I}(i)\right] = p_{_k}$, o estimador de Horwitz-Thompson pode ser reformulado como uma soma envolvendo apenas os n elementos da amostra:

$$\hat{\mu} = \frac{1}{N} \sum_{k=1}^{n} \frac{y_{_k}}{p_{_k}} = \frac{1}{N} \sum_{k=1}^{n} w_{_k} y_{_k}. \tag{2.2}$$

A variância amostral do estimador de Horwitz-Thompson é

$$\operatorname{var}\left(\hat{\mu}\right) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{\pi_{_{ij}}}{\pi_{_i} \pi_{_j}} - 1 \right) Y_{_i} Y_{_j}$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{_{ij}}}{\pi_{_i} \pi_{_j}} Y_{_i} Y_{_j} - \mu^2.$$

No caso de planos amostrais em que a amostra tem dimensão fixa,

$$\operatorname{var}(\hat{\mu}) = \frac{1}{2\,N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\pi_{_{ij}} - \pi_{_i}\pi_{_j}\right) \left(\frac{Y_{_i}}{\pi_{_i}} - \frac{Y_{_j}}{\pi_{_j}}\right)^2, \tag{2.3}$$

porque em amostras em que a dimensão é fixa, $\displaystyle\sum_{j=1}^{N} \left(\pi_{_{ij}} - \pi_{_i}\pi_{_j}\right) = 0$.

A expressão (2.3) mostra que em planos amostrais em que as probabilidades de inclusão $\pi_{_i}$ são proporcionais à magnitude de $Y_{_i}$ a variância do estimador de Horwitz-Thompson é baixa.

No caso de $Y_{_i} = y_{_k}$ e $Y_{_j} = y_{_\ell}$ e o plano amostral em $\mathcal{P} \times \mathcal{P}$ é próprio — as probabilidades de inclusão conjunta de $(Y_{_i}, Y_{_j})$ são positivas —

$$\pi_{_{ij}} = \mathbb{P}\left[\mathcal{I}(i)\mathcal{I}(j)\right] = p_{_{k\ell}} > 0$$

o estimador de $\operatorname{var}(\hat{\mu})$ é

$$\widehat{\operatorname{var}(\hat{\mu})} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{\pi_{_{ij}}} \left(\frac{\pi_{_{ij}}}{\pi_{_i}\pi_{_j}} - 1\right) \mathcal{I}(i) Y_{_i} \mathcal{I}(j) Y_{_j}$$

ou seja

$$\widehat{\operatorname{var}(\hat{\mu})} = \frac{1}{N^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \left(\frac{1}{p_{_k} p_{_\ell}} - \frac{1}{p_{_{k\ell}}}\right) y_{_k} y_{_\ell}. \tag{2.4}$$

Indubitavelmente o estimador de Horwitz-Thompson tem uma universalidade que o torna atraente, no sentido em que é facilmente usável com qualquer plano amostral próprio (em $\mathcal{P} \times \mathcal{P}$). Mas, por outro lado, é fácil exibir algumas patologias pouco agradáveis deste estimador. Por exemplo, se $Y_{_i} = \mu$ e considerarmos um plano amostral com probabilidades de inclusão desiguais e dimensão amostral $n = 1$,

$$\hat{\mu} = \frac{\mu}{N p_{_1}}$$

e

$$\operatorname{var}(\hat{\mu}) = \frac{\mu^2}{N^2} \left(\sum_{k=1}^{N} 1_{_{p_k}} - 1\right) > 0$$

uma vez que com probabilidades de inclusão desiguais a média aritmética e a média harmónica são necessariamente diferentes.

Evidentemente, se na amostra figurarem elementos da população com probabilidades de inclusão muito baixas, a estimativa da média pode ser desmedidamente inflacionada, e o efeito de probabilidades de inclusão conjunta ainda mais baixas

leva a que o peso que esses pares têm na estimação da variância possa ter um papel semelhante ao que nos faz temer a presença de *outliers* em muitas áreas de análise de dados. É este, provavelmente, o maior óbice à utilização geral do estimador de Horwitz-Thompson e a consequente limitação do seu uso a planos amostrais em que os pesos de amostragem sejam equilibradamente controlados.

2.5 Não-Resposta e Rarefação

A não-resposta é um problema incontornável em amostragem. Há sempre quem não esteja acessível, não queira responder a algumas das perguntas ou à totalidade de um questionário, e se a resposta ao questionário é feita de forma não presencial, pedindo por exemplo que se devolva pelo correio (usando um sobrescrito fornecido, com selo pago), uma larga fracção dos inquiridos não responde. Apesar de tentar incluir-se incentivos diversos, a taxa de resposta muitas vezes não atinge os 10%.

Há assim uma certa dose de lirismo, quase delírio, na teoria que expusemos, e vamos recordar brevemente, a teoria sobre a dimensão que a amostra deve ter para as estimativas terem, com um elevado grau de probabilidade, a acríbia desejada. A forma simplificada, porventura em excesso, é multiplicar n pelo inverso da estimativa da taxa de resposta.

Inspirando-nos em Aleixo *et al.* (2007) vamos basear-nos no facto de a rarefação de Rényi (1956) e a filtragem geométrica de Kovalenko (1965) serem assintoticamente equivalentes (no sentido em que levam à mesma lei limite, mas não deixe de se observar que uma limitação severa a esta abordagem é não estar provado que as velocidades de convergência para essa lei limite são iguais nos dois procedimentos) para propor metodologias mais consistentes.

Note-se, no entanto, que como admitem um processo de aleatorização, as variâncias são maiores — recorde-se que se

$$\operatorname{var}(X) = \mathbb{E}\left[\operatorname{var}(X \mid Y)\right] + \operatorname{var}\left[\mathbb{E}(X \mid Y)\right],$$

isto é, numa definição hierárquica a variância obtém-se como soma do valor médio da variância condicional com a variância do valor médio condicional, um resultado a que adiante teremos que recorrer repetidamente —, o que é um óbice a ter em conta!

No entanto, a aleatorização é o caminho natural para se obter resultados cada vez mais realistas, e isso leva naturalmente a uma observação: devido à escala a variabilidade é mais "visível": numa laranja vista a dois ou três metros de distância não se notam as rugosidades que se observam à distância do nosso braço.

Generalidades

A determinação da dimensão da amostra necessária para obter a precisão que desejamos quando estimamos parâmetros populacionais é um tema chave na teoria da amostragem e suas aplicações como ferramenta metodológica nas ciências experimentais. Por exemplo, se o nosso objectivo é estimar a média populacional a partir da média amostral, de modo a que a amplitude do intervalo de confiança $(1-\alpha)\times 100\%$ seja limitado por B, a dimensão da amostra $n = n_{_G}$ é o menor inteiro maior do que

$$\frac{4z^{2}_{1-\frac{\alpha}{2}}\sigma^{2}}{B^{2}}$$

no caso da amostragem com reposição (independência), e

$$\frac{\nu}{1+\frac{(\nu-1)B^{2}}{4z^{2}_{1-\frac{\alpha}{2}}\sigma^{2}}}$$

na amostragem aleatória simples sem reposição (isto é, permutabilidade em vez de independência) a partir de uma população finita de dimensão ν; na prática, o σ^{2} desconhecido é substituído por uma estimativa s^{2}, e o uso de quantis gaussianos $z_{1-\frac{\alpha}{2}}$ é justificado pelo teorema limite central clássico no caso de independência, e sua extensão para parcelas permutáveis (Erdös e Rényi, 1959), quando a amostragem é feita sem reposição a partir de populações finitas. No entanto, em muitas situações de amostragem, algumas das unidades seleccionadas para a amostra aleatória acabam por originar não-resposta. A fracção de não-respostas pode ser bastante elevada, e em inquéritos com resposta a devolver por correio, por exemplo, o questionário é enviado a um vasto número de indivíduos — uma regra prática *ad hoc* é $\frac{n_{_G}}{\tilde{p}}$, onde $\tilde{p}$ é a percentagem prevista de formulários devolvidos, usualmente uma estimativa grosseira baseada em estudos similares e populações alvo —, uma vez que a experiência acumulada mostra que apenas uma pequena percentagem p deles devolverá os formulários.

Vamos considerar o caso da rarefação (filtragem aleatória) em que, cada unidade incluída na amostra na etapa de planeamento permanece nela com probabilidade p, ou sai dela com probabilidade $1-p$, independentemente de qualquer outra.

Adiante investigamos os resultados obtidos ao usar uma amostra com dimensão aleatória $N \frown BinomialNegativa(n_{_G}, p)$, em vez da regra *ad hoc* de dimensão $\frac{n_{_G}}{\tilde{p}}$.

Para valores muito pequenos de p, o limite desse processo de filtragem é um processo de rarefação de Rényi (1956) da amostra inicial, e no ponto de vista dos resultados de Kovalenko (1965) e de Kozubovsky (1994), o processo de rarefação de

Rényi é equivalente a parar aleatoriamente a soma de variáveis aleatórias i.i.d., com subordinador independente $V \frown \text{Geométrica}(p)$.

A distribuição assintótica de $T = \sum_{k=1}^{V} X_k$, assumindo a existência da variância da distribuição parente, é Laplace generalizada e, em particular, Exponencial quando as parcelas são positivas. Este pode ser um resultado muito útil quando se faz amostragem de acontecimentos raros.

Note-se, no entanto, que o trabalho de Aleixo *et al.* (2007) capitaliza no facto de a rarefação de Rényi (1956) e as somas aleatórias com subordinadora geométrica de Kovalenko (1965) terem o mesmo limite, mas não discute uma questão relevante do ponto de vista prático quando se pretende obter o limite de uma das situações recorrendo à outra: será a velocidade de convergência para a lei limite a mesma? — é uma questão que, tanto quanto sabemos, se mantém em aberto.

O Método Delta

O método delta é, em traços muito gerais, a truncatura da expansão em série de Taylor de uma função com vista à obtenção de aproximações, nomeadamente, para os momentos de uma estatística de interesse. A prática habitual consiste em truncar a expansão após o primeira derivada. No caso de esta se anular no ponto em que é efectuada a expansão recorre-se ao termo correspondente à segunda derivada.

Suponhamos que em vez de um parâmetro θ desejamos estimar uma função desse parâmetro — por exemplo, queremos em estimar $\frac{1}{\mu}$ em vez de estimar o valor médio populacional μ. Mas o inverso de um variável aleatória com valor médio finito até podenão ter valor médio. Porém em situações regulares o recurso à expansão da função de interesse em série de Taylor pode proporcionar aproximações interessantes, truncando no termo linear ou no termo de segunda ordem. Vejamos:

Sejam $\boldsymbol{T} = (T_1, \ldots, T_n)$ e $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, onde $\theta_k = \mathbb{E}(T_k)$, $k = 1, \ldots, n$. Seja $g(\boldsymbol{T})$ um estimador de um parâmetro que nos interessa, sendo g uma função diferenciável, e denote-se

$$g'_k(\boldsymbol{\theta}) = \frac{\partial}{\partial t_k} g(\boldsymbol{t}) \mid_{\substack{t_1=\theta_1 \\ \cdots \\ t_n=\theta_n}} .$$

A expansão em série de Taylor de g em torno de $\boldsymbol{\theta}$, de primeira ordem, é

$$g(\boldsymbol{t}) = g(\boldsymbol{\theta}) + \sum_{k=1}^{n} g'_k(\boldsymbol{\theta})\,(t_k - \theta_k) + R_1.$$

Tomando valores médios,

$$\mathbb{E}\left[g(\boldsymbol{T})\right] \approx g(\boldsymbol{\theta}),$$

pois $\sum_{k=1}^{n} g'_k(\boldsymbol{\theta})\,\mathbb{E}\left(T_k - \theta_k\right) = 0$. Por outro lado, no que refere a variância do estimador,

$$\operatorname{var}\left[g(\boldsymbol{T})\right] \approx \mathbb{E}\left[\left(g(\boldsymbol{T}) - g(\boldsymbol{\theta})\right)^2\right] \approx \mathbb{E}\left[\left(\sum_{k=1}^{n} g'_k(\boldsymbol{\theta})\left(T_k - \theta_k\right)\right)^2\right]$$

obtendo-se a aproximação

$$\operatorname{var}\left[g(\boldsymbol{T})\right] \approx \sum_{k=1}^{n}\left[g'_k(\boldsymbol{\theta})\right]^2 \operatorname{var}\left(T_k\right) + 2\sum_{i>j} g'_i(\boldsymbol{\theta})\, g'_j(\boldsymbol{\theta}) \operatorname{cov}\left(T_i, T_j\right).$$

Pestana e Velosa (2010, pp. 1046–1048) enunciam os resultados fundamentais; Chandra (1999) contém uma exposição mais circunstanciada. Quer Diamantino (2008), quer Martins (2009), são complementos de leitura recomendáveis, uma vez que a exposição do método delta é ilustrada com um grande número de aplicações interessantes.

Estimando a Média

Considerem-se as duas situações seguintes:

1. $(X_1, \ldots, X_n)$ é uma amostra aleatória de dimensão n, onde os X_i independentes são tais que $X_i \stackrel{\mathrm{d}}{=} X$, com $\mathbb{E}(X) = \mu$ e $\operatorname{var}(X) = \sigma^2$. Nesta situação, podemos usar $\overline{X} = \dfrac{1}{n}\sum_{k=1}^{n} X_k$ para estimar μ com um limite de erro (erro padrão) B, com confiança $(1-\alpha)\times 100\%$, usando uma amostra de dimensão n_G, o menor inteiro maior do que $\dfrac{4 z^2_{1-\frac{\alpha}{2}}\,\sigma^2}{B^2}$.

2. $(X_1, \ldots, X_\nu)$ é uma população finita com média $\mu = \dfrac{1}{\nu}\sum_{k=1}^{\nu} X_k$ e variância $\sigma^2 = \dfrac{1}{\nu-1}\sum_{k=1}^{\nu}\left(X_k - \mu\right)^2$. A amostragem aleatória simples sem reposição garante que todas as $\binom{\nu}{n}$ amostras de dimensão n são equiprováveis. Note que, nesta situação, os X_k já não são independentes, mas a sua dependência mútua é fraca, e o teorema limite central para variáveis aleatórias permutáveis pode ser usado para determinar a dimensão da amostra necessária para obter a

precisão que desejamos: $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ estima μ com um limite de erro padrão B, com confiança $(1-\alpha)\times 100\%$, usando uma amostra de dimensão n_G, o menor inteiro maior do que $\dfrac{\nu}{1+\frac{(\nu-1)B^2}{4z^2_{1-\frac{\alpha}{2}}\sigma^2}}$. (Como não pode surgir confusão, usamos o mesmo símbolo n_G quer no caso independente, quer no caso permutável).

Assuma-se, no entanto, que sabemos que a amostra será sujeita a uma filtragem-p, isto é, cada X_k será efectivamente observado com probabilidade p, independentemente de cada um dos outros. Precisamos, portanto, de uma amostra maior de dimensão N, de modo a que a amostra filtrada tenha aproximadamente dimensão n_G. No que se segue iremos comparar os resultados usando uma amostra aleatória de dimensão $N \frown BinomialNegativa(n_G, p)$ com os resultados obtidos usando dimensão determinística $\dfrac{n_G}{p}$.

Observe que se $(Y_1, \ldots, Y_N)$ é uma amostra de Y_k independentes tais que $Y_k \overset{\mathrm{d}}{=} Y \frown Bernoulli(p)$, independente de $(X_1, \ldots, X_N)$, a amostra $(Z_1, \ldots, Z_{n^*})$ onde os Z_k são os $X_k Y_k$ não nulos, é uma amostra filtrada-p, e $T = \sum_{i=1}^{n^*} Z_i = \sum_{k=1}^{N} X_k Y_k$. Observe que $\mathbb{E}(X_k Y_k) = p\mu$ e que $\mathrm{var}(X_k Y_k) = p(\sigma^2 + (1-p)\mu^2)$.

Dimensão da amostra determinística usando a regra *ad hoc* — estimador $\tilde{\mu}_1 = \dfrac{T}{n_G}$

Se usarmos $N = \frac{n_G}{p}$, o valor esperado e a variância do estimador $\tilde{\mu}_1 = \frac{T}{n_G}$ são:

1. No caso da amostragem independente,

$$\mathbb{E}(\tilde{\mu}_1) = \mu \qquad \text{e} \qquad \mathrm{var}(\tilde{\mu}_1) = \frac{\sigma^2 + (1-p)\mu^2}{n_G}.$$

Portanto, a precisão da estimativa será muito pior do que a esperada sempre que $\mu \gg 0$.

2. Na amostragem aleatória simples sem repetição a partir de uma população com dimensão ν, com a correcção da dimensão da amostra finita para a variância, obtemos resultados semelhantes:

$$\mathbb{E}(\tilde{\mu}_1) = \mu \qquad \text{e} \qquad \mathrm{var}(\tilde{\mu}_1) = \frac{\nu - N}{\nu - 1}\,\frac{\sigma^2 + (1-p)\mu^2}{n_G},$$

e, portanto, há um valor $\frac{\nu - N}{\nu - 1}\frac{(1-p)\mu^2}{n_G}$ a mais na variância do estimador, quando comparado com a situação de não filtragem.

Dimensão da amostra aleatória Binomial Negativa — estimador $\tilde{\mu}_2 = \frac{T}{pN}$

Seja $N \frown BinomialNegativa(n_G, p)$, $\mathbb{E}(N) = \frac{n_G}{p}$, $\text{var}(N) = \frac{n_G(1-p)}{p^2}$, e $\mathbb{E}(N^2) = \frac{n_G(n_G+1-p)}{p^2}$, e consideremos o estimador $\tilde{\mu}_2 = \frac{T}{pN}$.

Usando a expansão de Taylor linearmente truncada,

$$\tilde{\mu}_2 = \frac{T}{pN} \approx \frac{\mathbb{E}(T)}{p\mathbb{E}(N)} + \frac{1}{p\mathbb{E}(N)}\left[T - \mathbb{E}(T)\right] - \frac{\mathbb{E}(T)}{p[\mathbb{E}(N)]^2}\left[N - \mathbb{E}(N)\right].$$

$\mathbb{E}(T) = \mathbb{E}\left[\mathbb{E}(T|N)\right] = \mathbb{E}(Np\mu) = \frac{n_G}{p}p\mu = n_G\mu$, e assim $\mathbb{E}(\tilde{\mu}_2) \approx \frac{n_G\mu}{p\frac{n_G}{p}} = \mu$. No que respeita à variância do estimador $\tilde{\mu}_2$:

1. No caso da amostragem independente, $\text{var}(T|N) = Np(\sigma^2 + (1-p)\mu^2)$, e assim

$$\text{var}(T) = \mathbb{E}\left[\text{var}(T|N)\right] + \text{var}\left[\mathbb{E}(T|N)\right] = n_G(\sigma^2 + 2(1-p)\mu^2).$$

Portanto,

$$\text{var}(\tilde{\mu}_2) \approx \frac{1}{n_G^2}n_G(\sigma^2 + 2(1-p)\mu^2) + \left(\frac{n_G\mu}{p\left(\frac{n_G}{p}\right)^2}\right)^2 \frac{n_G(1-p)}{p^2} = \frac{\sigma^2 + 3(1-p)\mu^2}{n_G}.$$

2. Na população finita, com dimensão ν, temos o factor de correcção para populações finitas, $\text{var}(T|N) = \frac{\nu-N}{\nu-1}Np(\sigma^2 + (1-p)\mu^2)$, e assim

$$\text{var}(T) = n_G\left[\frac{(\nu+1)p - (n_G+1)}{p(\nu-1)}\right]\sigma^2 + n_G\left[1 + \frac{(\nu+1)p-(n_G+1)}{p(\nu-1)}\right](1-p)\mu^2.$$

Portanto

$$\text{var}(\tilde{\mu}_2) \approx \frac{\left[\frac{(\nu+1)p-(n_G+1)}{p(\nu-1)}\right]\sigma^2 + \left[2 + \frac{(\nu+1)p-(n_G+1)}{p(\nu-1)}\right](1-p)\mu^2}{n_G}.$$

Em consequência, $\tilde{\mu}_2$ é menos eficiente do que $\tilde{\mu}_1$. Observe que o valor esperado do denominador de $\tilde{\mu}_2$ é n_G e, portanto, estamos a dividir uma soma mais variável T (com um número aleatório de parcelas) por um valor próximo do mesmo n_G que usámos no denominador de $\tilde{\mu}_1$, e assim este resultado faz sentido.

Dimensão da amostra aleatória Binomial Negativa — estimador $\tilde{\mu}_3 = \dfrac{T}{W}$

Uma abordagem mais sofisticada seria contar o número de parcelas não nulas, $W = \sum_{k=1}^{N} Y_k$. Sendo $W|N \frown Binomial(N,p)$, para dividir a soma $T = \sum_{k=1}^{N} X_k Y_k$, isto é, para usar o e stimador $\tilde{\mu}_3 = \dfrac{T}{W}$.

No entanto, neste cenário $\mathbb{E}(W) = n_G$, $\text{var}(W) = 2n_G(1-p)$ e a variância de $\dfrac{T}{W} \approx \mu + \dfrac{1}{n_G}(T - n_G\mu) - \dfrac{\mu}{n_G}(W - n_G)$ é

1. No esquema de amostragem independente

$$\text{var}(\tilde{\mu}_3) \approx \frac{1}{n_G^2} n_G(\sigma^2 + 2(1-p)\mu^2) + \left(\frac{\mu}{n_G}\right)^2 2n_G(1-p) = \frac{\sigma^2 + 4(1-p)\mu^2}{n_G}.$$

2. Na população finita de dimensão ν, a variância de $\tilde{\mu}_3$ quando a amostragem é sem reposição é

$$\text{var}(\tilde{\mu}_3) \approx \frac{\left[\dfrac{(\nu+1)p - (n_G+1)}{p(\nu-1)}\right]\sigma^2 + \left[3 + \dfrac{(\nu+1)p - (n_G+1)}{p(\nu-1)}\right](1-p)\mu^2}{n_G}.$$

Portanto, quanto mais variabilidade introduzimos, menor eficiência obtemos do estimador. Parece que a única forma de alcançar os nossos objectivos seria usar um esquema de amostragem inversa desajeitado, continuando a amostragem até a dimensão da amostra atingir n_G.

Nas secções seguintes tentaremos uma abordagem alternativa, quando o parâmetro de filtragem p está próximo de zero. Esta filtragem radical foi designada "rarefação" por Rényi (1956).

Filtragem Geométrica

Seja $Y = \sum_{k=1}^{V} X_k$, onde os X_k variáveis aleatórias independentes tais que as parcelas $X_k \stackrel{\mathrm{d}}{=} X \geq 0$, $k = 1, 2, \ldots$, são independentes da variável subordinadora $V \frown$ Geométrica(p_n), e $\mathbb{E}(X) = \mu$. Consequentemente a média da soma aleatoriamente interrompida geométrica Y é $\delta = \dfrac{\mu}{p_n}$. A função característica de Y é $\varphi_Y(t) = \mathcal{G}_V(\varphi_X(t))$, onde $\mathcal{G}_V$ é a função geradora de probabilidade de V. Então

$$\varphi_Y(p_n t) = \frac{1}{1 + \dfrac{1 - \varphi_X(p_n t)}{p_n \varphi_X(p_n t)}} = \frac{1}{1 + \dfrac{1 - \varphi_X(p_n t)}{p_n t} \dfrac{t}{\varphi_X(p_n t)}}.$$

Como $\dfrac{1 - \varphi_X(p_n t)}{p_n t} \xrightarrow[p_n \to 0]{} -\varphi'_X(0) = -\mathrm{i}\mu$ e $\varphi_X(p_n t) \xrightarrow[p_n \to 0]{} \varphi_X(0) = 1$, o limite do lado direito em cima é $\frac{1}{1 - \mathrm{i}\mu t}$ e consequentemente $\varphi_Y(t) = \dfrac{1}{1 - \mathrm{i}\delta t}$, que é a função característica de uma variável aleatória exponencial com média δ.

Assim, o limite de um processo geometricamente rarefeito com parcelas positivas com valor esperado finito é exponencial. Este resultado assintótico para o processo de rarefação foi inicialmente descoberto por Rényi (1956); Kovalenko (1965) estabeleceu que as transformadas de Laplace de variáveis aleatórias positivas que são estáveis com respeito à rarefação elementar são da forma $L(s) = \dfrac{1}{1 + cs^{\delta}}$, $c > 0$, $\delta \in (0, 1]$, o caso $\delta = 1$ — isto é, limite exponencial — correspondendo a variância finita. Os resultados de Kovalenko mostram que isto coincide com a classe de somas aleatóriamente interrompidas $\sum_{k=0}^{V} X_k$ com parcelas positivas i.i.d., independentes da variável subordinadora $V \frown$ Geométrica(p). A teoria geral das somas aleatoriamente interrompidas geométricas (Kozubowski, 1994) conduz a resultados semelhantes para a função característica de parcelas cujo suporte não é necessariamente positivo, e em particular a lei limite simétrica para somas geométricas de parcelas independentes de variância finita é a distribuição Laplace.

Não conhecemos qualquer investigação sistemática sobre até onde a rarefação deve ir de modo a que o resultado assintótico possa ser tomado como uma boa aproximação. A exponencial é estável relativamente à filtragem-p, ou seja, a exponencial filtrada-p é ainda exponencial, qualquer que seja o valor de $p \in (0, 1]$, no entanto, isto é uma situação excepcional.

Assuma-se agora que $Y = \sum_{k=1}^{V} X_k \overset{\circ}{\frown} Exponencial(\delta)$; de

$$W = \frac{Y}{V} \approx \mu + p(Y - \delta) - p^2\delta\left(V - \frac{1}{p}\right)$$

obtemos que $\mathbb{E}(W) \approx \mu$ e $\text{var}(W) \approx (2-p)\mu^2$.

Sejam $(W_1, \ldots, W_n)$ réplicas independentes de W, $\tilde{\mu}_4 = \overline{W} = \frac{1}{n}\sum_{k=1}^{n} W_k$, para as quais o teorema do limite central se verifica. Como a variância de $\tilde{\mu}_4$ é $\frac{(2-p)\mu^2}{n}$, se pretendermos que o intervalo de confiança a $(1-\alpha)\times 100\%$ seja limitado por B, devemos tomar uma amostra de dimensão n_E, o menor inteiro maior do que $\frac{4z^2_{1-\frac{\alpha}{2}}(2-p)\mu^2}{B^2}$.

Remetemos para mais detalhes para Diamantino (2008), que usou uma interessante família de distribuições (que no caso assimétrico positivo generaliza as Weibull, e no caso simétrico contém a Laplace e a Gaussiana, e representa muitas situações de curtose) para comparar $\tilde{\mu}_4$ com $\tilde{\mu}_1$.

Capítulo 3

Um Exemplo e Algumas Reflexões Conclusivas

3.1 Erros Alheios à Amostragem

Porventura devido a uma credulidade ingénua no *slogan* "Informação é poder", há uma grande avidez na aquisição de informação, através de sondagens de qualidade variável.

Está por provar que informação, qualquer, traga poder ou mesmo vantagens; não há dúvida de que informação estratégica é um bom trunfo quando é necessário tomar decisões sob incerteza.

A par da vertente matemática da Teoria da Amostragem, de que expusemos no Capítulo 2 alguns aspectos, há muitas outras valências e ciências que é preciso conhecer para se obter informação de qualidade. E, sobretudo, o bom senso é imprescindível. A este propósito, não podemos deixar de comentar que a profusão de sondagens, inquéritos, estudos de opinião, estudos de mercado, operações de *marketing* e outras, que massacram o público, denota uma grande falta de bom senso colectivo, pois têm tido como resultado inevitável um declínio notório da taxa de respondentes, comprometendo a qualidade de sondagens que o INE e instituições oficiais em que delega competências têm que fazer para suporte de decisões importantes para o interesse público.

De facto, a vertente matemática da Teoria da Amostragem tem como objecto privilegiado de investigação o *erro amostral*, o erro decorrente da variabilidade das estimativas de amostra para amostra. Há no entanto outro tipo de erros, que de um modo geral só podem ser parcialmente controlados, e que nada têm que ver com a variabilidade inerente dos estimadores baseados em amostras aleatórias, erros esses em geral designados pela expressão "*erros alheios à amostragem*".

Tanur (2011) classifica-os em cinco grupos: erros de especificação, erros de cober-

tura, não-resposta, erros nas respostas, e erros de processamento. Esta classificação dos erros alheios à amostragem é a adoptada por Tanur (2011), e simplifica a classificação de Deming (1944), que arrolou 13 factores que afectam a utilidade das sondagens.

Este trabalho de Deming corresponde a um progresso notável, uma vez que anteriormente a ele Neyman (1934), por exemplo, assume ingenuamente que os dados de amostragem estão isentos de erros não amostrais. A partir do trabalho de Deming (1944) deixou de ser possível assumir que as sondagens usavam, por definição, dados sem erros, e o erro total (de que Biemer, 2010, apresenta uma panorâmica interessante) passou a ser uma área de investigação importante. Recomenda-se ainda a leitura de Mosteller (1978), que como Tanur (2011) muito justamente comenta estará ultrapassado nas técnicas, mas mantém-se actual — e profundo — nos conceitos.

Os **erros de especificação** ocorrem na fase de planeamento, e correspondem na generalidade a adoptar metodologias de recolha de informação que não são os adequados para a obtenção dos dados necessários para atingir os objectivos propostos. Por exemplo, não colocar algumas questões pertinentes — num estudo sobre sucesso escolar, não inquirir sobre a qualidade da alimentação ou sobre as horas de sono. Formular perguntas ambíguas (o uso da escala de Lickert proporciona abundantes exemplos — pedir que se classifique na escala de 1 a 10 a severidade da dor que o paciente sente é mais propício a confundir o nosso entendimento da realidade do que a esclarecê-lo). Colocar directamente questões embaraçosas, em vez de usar aleatorização das perguntas ("*random response*") é um convite à mentira ou recusa a responder. Perguntas tendenciosas, que influenciam as respostas dos inquiridos, são um erro de especificação comum, uma vez que as camadas sociais mais fragilizadas tentam responder por forma a causar boa impressão.

Adiante apresentamos duas versões de um inquérito sobre hábitos alimentares de estudantes; na essência, as perguntas são as mesmas nos dois inquéritos, mas num deles as perguntas foram viciadas ("*loaded*") com uma introdução sobre o que é uma refeição equilibrada e quais os nutrientes necessários para o cérebro ter um bom desempenho. Contrariamente ao que se esperava, os resultados dos dois inquéritos levam a concluir que não há heterogeneidade de duas populações. De facto trata-se de uma única população (os formulários foram intercalados por forma a em cada turma se entregarem alternadamente uma e outra redacção), que deve estar suficientemente condicionada para resistir à viciação das perguntas.

Os **erros de cobertura** mais comuns são por defeito: parte das unidades da população alvo estão ausentes da base de amostragem que é usada. Um exemplo simples: um inquérito telefónico que só use telefones fixos deixaria actualmente muitos habitantes de Portugal fora da base de amostragem. Por outro lado, podem ocorrer também erros de cobertura por excesso, por exemplo quando os inquiridos não deveriam pertencer, de facto, à base de amostragem — por exemplo, quando o

entrevistador aproveita, para ser pago, as respostas de um menor numa sondagem sobre intenções de voto. Note-se que a escolha da base de amostragem pode ter uma influência grande em erros de interpretação sobre a estratégia amostral que está a ser usada. Por exemplo se se usar amostragem aleatória simples sem reposição na selecção de números de telemóvel, os inquiridos estão a ser seleccionados afinal de acordo com um plano amostral em que a probabilidade de selecção varia com o número de telemóveis de cada potencial inquirido, e deve-se na estimação do total, por exemplo, ter o cuidado de usar a teoria de Hansen and Hurwitz (1943).

A **não-resposta** é um problema maior no que se refere ao erro alheio à amostragem. Scheaffer *et al.* (1996) afirmam, "*Probably the most serious of all nonobservational errors, however, is nonresponse*". Por isso no Capítulo 2 nos preocupámos tanto com possíveis formas de recalcular o tamanho da amostra em função da taxa de não-resposta.

De facto, há sempre quem não esteja acessível, não queira responder a algumas das perguntas ou à totalidade de um questionário, e se a resposta ao questionário é feita de forma não presencial, pedindo por exemplo que se devolva pelo correio (usando um sobrescrito fornecido, com selo pago), uma larga fracção dos inquiridos não responde. Apesar de tentar incluir-se incentivos diversos, a taxa de resposta muitas vezes não atinge os 10% (Bourque and Fielder, 2003, p. 16). Diversos livros justamente célebres em amostragem (Scheaffer *et al.*, 1996); Barnett, 2002; Singh, 2003, para citar apenas alguns) dedicam algumas páginas ou mesmo capítulos inteiros a não-resposta e como acomodá-la, e é interessante observar as estimativas cada vez mais pessimistas sobre a taxa de não-resposta com a banalização das sondagens. Bethlehem (2011) considera que em geral excede 50%, e muitos estudos de casos mostram que pode rondar os 90%, e exceder mesmo esse valor no caso de não se estimular o retorno de respostas, ou o inquérito ser de alguma forma incómodo, ou simplesmente maçador.

Barnett (2002) dá indicações úteis sobre como contornar o viés devido a não-resposta usando pós-estratificação. Mas esse método admite que se sabe quem não respondeu. Ora nos muito difundidos inquéritos usando as facilidades da Web nem a taxa de resposta é conhecida! O interessante verbete de Manfreda *et al.* (2011) levanta questões pertinentes, e possivelmente a comunidade estatística, e nomeadamente a *IASS — International Association of Survey Statisticians*, deveriam não só investir mais na investigação da qualidade das sondagens usando a *Web* e alertar os utilizadores deste tipo de sondagens do viés que elas podem ocasionar. Interrogamo-nos, por exemplo, sobre o valor das conclusões do *Estudo da Satisfação e Motivação dos Académicos no Ensino Superior Português (ESMAESP) - PTDC/ESC/67784/2006* a que respondemos recentemente, pois é tipicamente uma questão em que os grupos de respondentes e de não-respondentes podem diferir muito susbtancialmente entre si.

Singh (2003) tem um tratamento em profundidade do problema da não-resposta,

e do tratamento de dados omissos, que é também uma consequência da não-resposta. É a nossa recomendação para um primeiro contacto com o importante tema de imputação de dados omissos (inclusive imputação múltipla), um tema de investigação em que muito se tem evoluído mas que continua a ser um desafio a nível de estatísticas oficiais, apesar de a liberalização progressiva do recurso a dados administrativos ter sido nos últimos anos um paliativo no tratamento dessa questão tão problemática.

Os **erros nas respostas** devem-se a múltiplos factores, nomeadamente a serem colocadas perguntas ambíguas, ou embaraçosas, ou a que o inquirido não sabe responder, ou que acumulam estes problemas. Por exemplo, no Censo 2001 o INE foi forçado a incluir uma pergunta sobre o número de deficientes que viviam na habitação — deficiência é um conceito vago para a generalidade dos inquiridos e dos agentes recenseadores, com interpretações muito diversas. Os inquéritos muito longos, e/ou de resposta obrigatória, levam muitas vezes os inquiridos a dar respostas fáceis falsas. No estudo de prevalência de doenças, os inquiridos podem ignorar que são portadores, ou terem vergonha de admitir que têm a doença em causa, se por exemplo puder ser adquirida por transmissão sexual, ou falta de higiene.

Os **erros de processamento** podem ocorrer em diversas fases do estudo. Inclusivamente, podem começar logo na fase de notação da informação. É célebre a frase de Sir Josiah Stamp "*Public agencies are very keen on amassing statistics — they collect them, add them, raise them to the n-th power, take the cube root and prepare wonderful diagrams. But what you must never forget is that every one of those figures comes in the first instance from the village watchman, who just puts down what he damn well pleases.* "

3.2 Questões que Ajudam a Construir um Inquérito

O arrolamento acima feito demonstra a necessidade de investir tempo e esforço na preparação de uma operação de sondagem. Como em todas as coisas que queremos fazer bem, devemos planear tendo em conta as questões fundamentais, nomeadamente

- para quê? (e porquê?)
- o quê
- a quem?
- como?
- quem?

A primeira questão é essencial para a definição dos objectivos e, conjuntamente com a segunda questão, para limitar os erros de especificação. A terceira questão contribui para evitar erros de cobertura.

A quarta questão tem que ver com não-resposta e com erros nas respostas, e é uma questão muito complexa. A repetição desta questão em várias fases levou a que o inquérito que adiante apresentamos tenha demorado alguns meses a ser burilado, passando por mais do que uma dezena de versões.

A quinta questão tem como objectivo pensar na formação dos entrevistadores, agentes recenseadores e outros, que têm intervenção directa nos erros de processamento.

Uma vez que nos ocupámos sobretudo da colecção de informação por inquérito auto-administrado, referimos alguns pontos em que tivemos que reflectir.

Em operações de recolha de informação, os inquéritos têm evidentes vantagens: são eficientes na recolha de informação em que podem ser usadas técnicas estatísticas para determinar a validade, a fiabilidade e a significância dos resultados; são flexíveis, no sentido em que pode ser recolhida uma grande variedade de informação, porque se adaptam igualmente bem a investigar atitudes, valores , comportamentos, opiniões, etc; têm em geral uma boa relação qualidade / preço, devido à focalização providenciada por questões padronizadas.

Mas por outro lado têm desvantagens que devem ser ponderadas antes de se optar por tal instrumento de recolha: a honestidade e a capacidade de compreensão para responder às questões que são colocadas é um factor essencial; não dão bons resultados no estudo de fenómenos sociais complexos; a amostra tem que ser representativa, caso contrário a inferência sobre os parâmetros populacionais é enviesada.

Para usufruir das vantagens indicadas, qualquer inquérito deve ter as seguintes características:

1. ser claro nos seus objectivos;
2. usar linguagem acessível a todos os inquiridos;
3. as respostas devem traduzir a opinião dos inquiridos, pelo que há que ter muito cuidado em não viciar as perguntas;
4. os dados devem ser sujeitos a análise estatística de forma a inferir os resultados e posteriormente tomar decisões;
5. O intervalo entre o tempo do seu planeamento e a obtenção de resultados deve ser breve, pois podem ocorrer alterações na população que retirem qualidade ao questionário;
6. As questões devem focar-se em ocorrências recentes, sob perigo de a memória alterar a verdade (efeito telescópico).

Anote-se ainda que não só a formulação das perguntas como também o seu sequenciamento podem viciar um inquérito. A fase de construção do questionário é, de facto, crucial para a obtenção de boa informação.

Numa sondagem tanto nos podem interessar variáveis quantitativas como qualitativas, sendo estas últimas mais fáceis de recolher. O instrumento de notação, sobretudo em questionários auto-administrados, é de grande importância, e os profissionais manifestam em geral preferência pelo que se designa por "resposta fechada" — que evidentemente tem o perigo de proporcionar uma mera caricatura redutora da realidade.

Um erro típico de inquéritos que, apesar de auto-administrados têm mecanismos concebidos para evitar não-resposta (por exemplo inquéritos concebidos para serem respondidos na *Web*), é o conjunto de respostas previstas nas perguntas fechadas não ser um arrolamento exaustivo e disjunto (partição) do universo de possíveis respostas. Confrontado com essa situação, o inquirido mente, ou simplesmente abandona o inquérito, e em vez de haver uma não-resposta a um item, passa a haver não-resposta global. Mas como já atrás referimos, a questão da não-resposta é particularmente espinhosa em inquéritos em que é impossível moderar o enviesamento decorrente de não-resposta usando pós-estratificação ou imputação múltipla.

3.3 Alimentação e o Sucesso Escolar

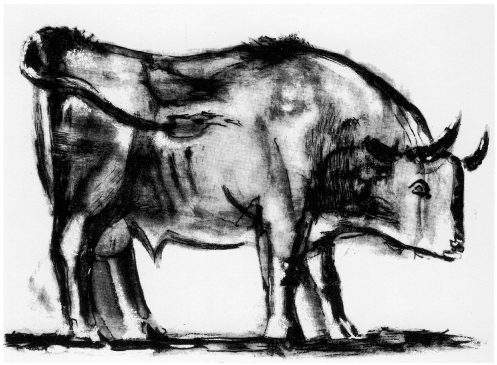
A fim de elucidar com um exemplo concreto algumas das questões atrás referidas, elaborámos um inquérito sobre alimentação e o sucesso escolar, avaliado numa perspectiva subjectiva pelo próprio aluno.

O inquérito tem quatro versões, duas para aplicar na FCUL, duas para aplicar num estabelecimento de ensino secundário, particular, em que o autor é docente. Em cada instituição um dos inquéritos tem as perguntas apresentadas de forma neutra, sem qualquer comentário tendente a influenciar as respostas dos alunos, enquanto na outra versão algumas das questões estão precedidas de um preâmbulo que sugere que há uma resposta mais "inteligente".

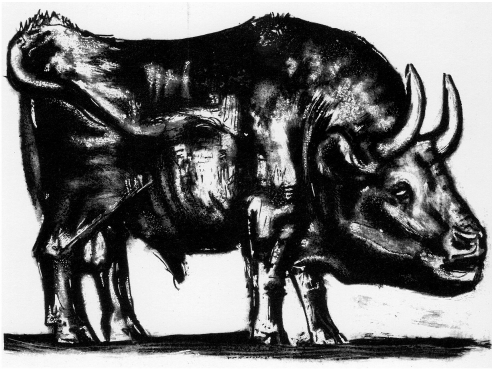
Em cada estabelecimento de ensino, houve o trabalho prévio de intercalar inquéritos com formulação neutra e inquéritos com formulação viciada das questões. Assim, obtiveram-se amostras equilibradas, e deixámos o acaso decidir quem respondia a quê, uma forma de aleatorização vulgarmente usada pela sua simplicidade de implementação.

As diferentes versões estão reproduzidas em apêndice. Gostaríamos de salientar que a elaboração deste inquérito, para se chegar à aparência final simples que tem, demorou algumas semanas de trabalho de aperfeiçoamento incorporando as correcções sugeridas pelas críticas dos meus orientadores. Uma colecção de provas

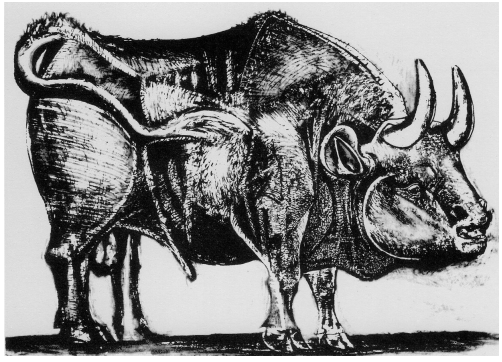
de artista de Picasso[(1)], para fazer a gravura de um touro, é uma boa metáfora do trabalho de preparação de um inquérito: numa fase inicial foram incluídos muitos pormenores, o trabalho de desbastar, simplificar, chegar à essência, é o fundamental para se chegar a uma fase definitiva com a qualidade pretendida.



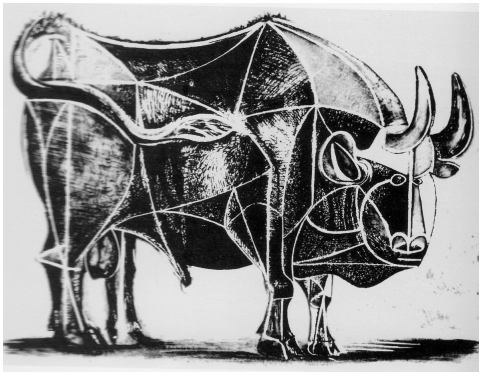
Versão 1
$1945 - 12 - 05$



Versão 2
$1945 - 12 - 12$

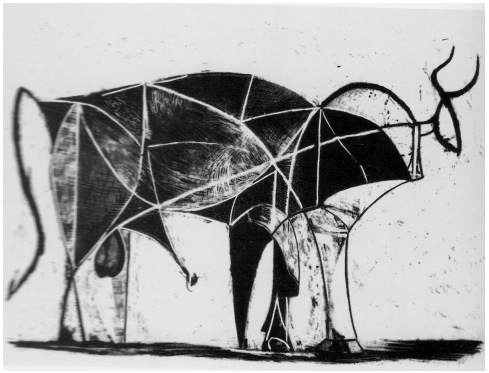


Versão 3
$1945 - 12 - 18$

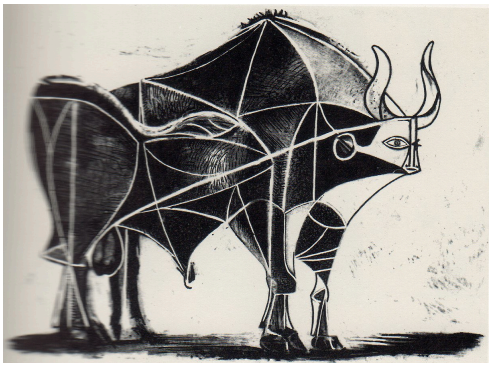


Versão 4
$1945 - 12 - 22$

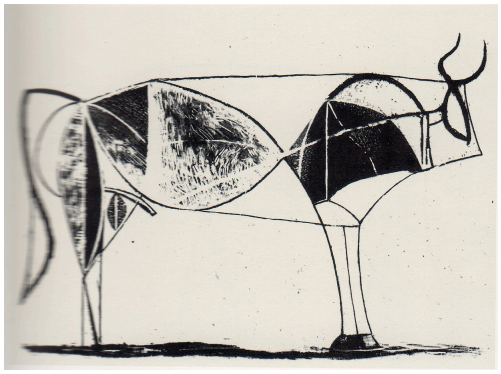
(1) F. Mourlot (1970). *Picasso Litographe*, André Sauret Éditions du Livre, Paris, pp. 27–30. Pelas datas das versões observa-se que Picasso foi mais rápido a alcançar o resultado final, o que decerto se explica por não ter um horário sobrecarregado num colégio, ou então ter aproveitado bem as férias de Natal.



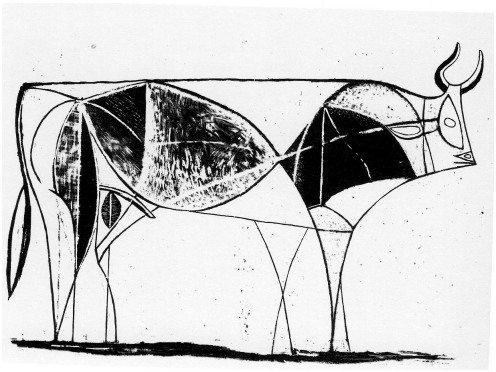
Versão 5
$1945 - 12 - 24$



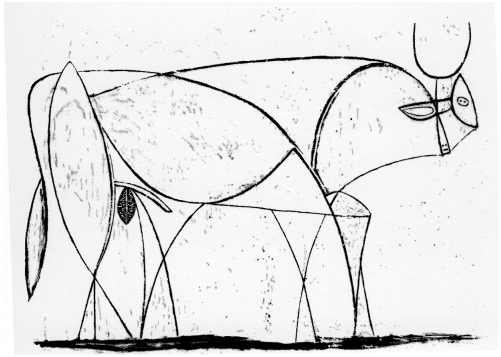
Versão 6
$1945 - 12 - 26$



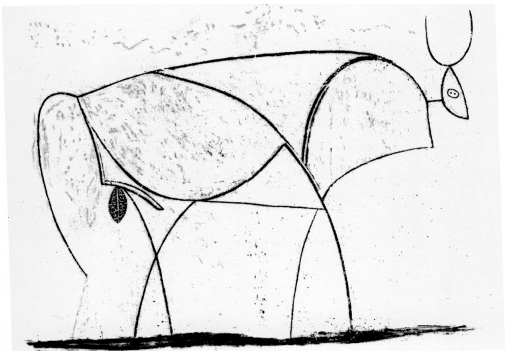
Versão 7
$1945 - 12 - 28$



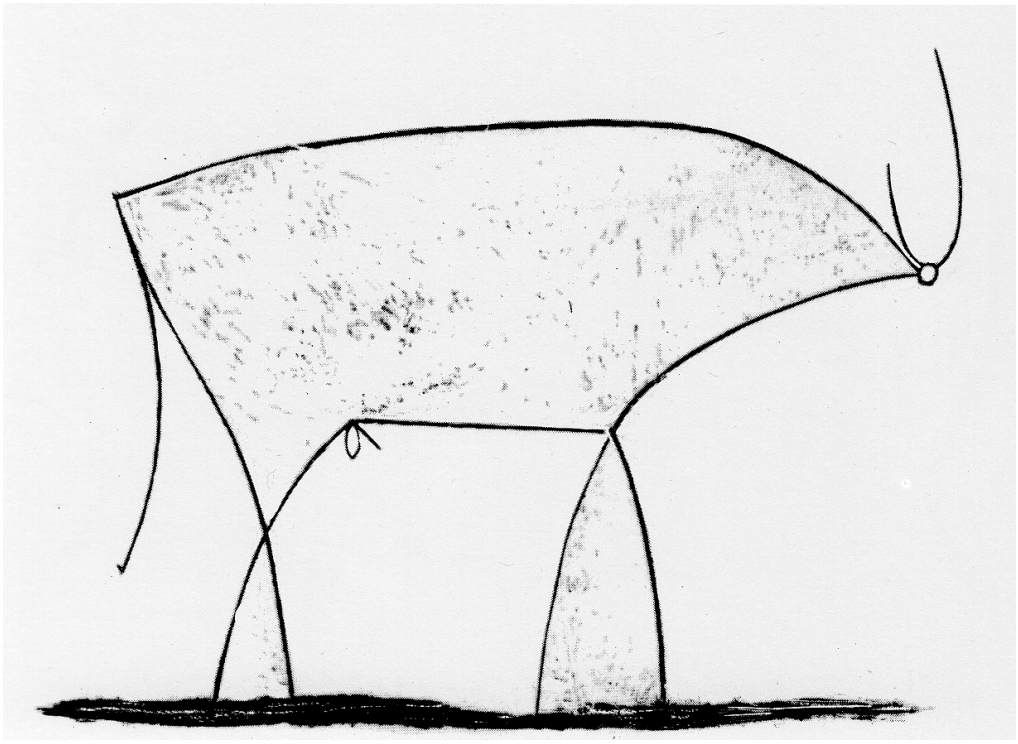
Versão 8
$1946 - 01 - 02$



Versão 9
$1946 - 01 - 05$



Versão 10
$1946 - 01 - 10$



Versão final
$1946 - 01 - 17$

Partindo de uma versão com tantas perguntas e pormenores, enchendo duas dezenas de páginas, a que possivelmente só três inquiridos iriam fastidiosamente responder, impacientando-se com a minúcia das questões e provavelmente respondendo com irritação a partir da quinta folha o que primeiro lhes viesse à cabeça, fomos podando até chegar a uma versão simples, adequada para investigar o que tínhamos em vista — e que era obter um estudo piloto, por um lado, e por outro lado ter um exemplo efectivo que nos permitisse quantificar o efeito das perguntas viciadas nas percentagens de respostas "politicamente correctas". Pareceu-nos também interessante comparar o efeito das questões viciadas em duas populações, diferentes em classe etária e em *status* social, para observar se se notava maior resistência às perguntas viciadas na população mais amadurecida.

Apontamos um defeito: na apresentação do inquérito aos inquiridos, não consta uma estimativa do tempo que gastarão a responder. Teria sido melhor haver logo uma indicação de que responder a este inquérito iria tomar 10 a 15 minutos de tempo — uma estimativa conservadora, e que por isso mesmo evita protestos "afinal demora mais tempo do que diziam!".

Breve Sumário das Conclusões

No CD que acompanha esta dissertação encontram-se quatro ficheiros .sav com a codificação das respostas:

- Univ[A].sav contém as 80 respostas de alunos da FCUL ao inquérito do Apêndice A;
- Univ[B].sav contém as 70 respostas de alunos da FCUL ao inquérito do Apêndice B
- Sec[c].sav contém as 65 respostas ao inquérito reproduzido no apêndice C, aplicado a alunos do ensino secundário;
- Sec[D].sav contém as 41 respostas ao inquérito reproduzido no apêndice D, aplicado a alunos do ensino secundário.

A fim de não limitar o acesso apenas a quem pode utilizar SPSS, incluem-se também no CD os ficheiros UnivAA, UnivBB, SecCC, SecDD, que são a exportação dos anteriormente referidos para EXCEL; os ficheiros UnivCompleto.xlsx e SecCompleto.xlsx contêm a concatenação de UnivAA e UnivBB, e de SecCC e SecDD, respectivamente.

Note-se que existem alguns erros de codificação (0.1 quando o resultado devia ser ou 0 ou 1, 10, em situação análoga), que não corrigimos exactamente para exemplificar que as "gralhas" podem escapar, mesmo na fase de análise dos dados, pois o computador executa cegamente os procedimentos que se lhe pede.

Uma análise rápida das respostas — pois não é obviamente objectivo de uma dissertação de mestrado fazer uma análise descritiva — mostra que quer no secundário quer na universidade, a colecção de dados (reconhecemos que não é uma amostra, com as boas qualidades discutidas anteriormente) é constituída maioritariamente por alunos que

- não gastam muito tempo em viagens entre a habitação e a escola,
- vivem com os pais,
- mais especificamente, vivem com ambos os progenitores,
- são suportados por mesada fornecida pelos pais,
- escolhem o local de refeições devido ao preço,
- não tomam refeições no estabelecimento de ensino,
- consomem produtos lácteos ao pequeno almoço,

- optam por refeições diferentes ao almoço e ao jantar.

No que se refere aos alunos da FCUL, a colecção de dados obtidos foi equilibrada no que se refere à razão masculino/feminino, enquanto no secundário cerca de 70% das resposta foi de alunos. As classes modais foram diferentes no que se refere às opções alimentares e bebidas consumidas ao almoço. A questão sobre gasto semanal com refeições teve respostas inadequadas, o que mostra que o inquérito é longo, e as respostas vão perdendo qualidade.

Como já afirmámos, não deve ser objectivo de uma dissertação uma análise descritiva ociosa. Tendo cometido o erro de principiante de anteceder algumas questões de frases que indicavam qual era a escolha "louvável", foi-nos pedido que usássemos inquéritos com esse erro e inquéritos em que as perguntas tinham a formulação neutra que devem ter, para observar a que ponto a formulação viciada influenciava as respostas. Tal poderia eventualmente ser observado nas respostas às questões 10, 11 e 13.

No que se refere à questão 10, por exemplo, considerámos que o tratamento adequado seria com a análise das tabelas de contingência para testar se havia homogeneidade de respostas aos inquéritos viciados e neutros.

No que se refere a alunos universitários, o valor observado da estatística $X^2_{3,2} = 0.099$ não corresponde a um valor de prova preocupante; o mesmo se pode dizer da análise da tabela correspondente aos alunos do secundário, $X^2_{3,2} = 2.482$, muito inferior ao quantil crítico $5.99 = \chi^2_{2,\,0.95}$.

No que se refere à questão 13, é evidente que não há diferenças entre respostas a inquéritos viciados e neutros na colecção de dados de alunos da FCUL (médias 3.7375 e 3.7428, respectivamente no grupo "viciado"e no grupo "neutro").

Já no que se refere à colecção de dados de alunos do secundário, no grupo neutro, com 65 observações, a média foi 3.8153 e o desvio padrão 0.429, e no grupo viciado, com 65 observações, a média foi 3.6585 e o desvio padrão 0.575; apesar de a variável não ser gaussiana, resolvemos usar como aproximação o teste t. Neste caso, a homocedasticidade não é rejeitada, o desvio padrão ponderado é 0.490, e o valor observado da estatística de teste é 1.604, a que corresponde um valor de prova $p = 0.058$.

Consequentemente, também neste caso não se pode concluir que viciar a forma de questionar influenciou de forma significativa as respostas.

Note-se que no ensino secundário as formas de alimentação correcta são amplamente discutidas, e muitas vezes objecto de trabalhos, o que porventura explica este resultado, contrário ao que esperávamos.

Apêndice A

Inquérito
(FCUL, questões neutras)

A ALIMENTAÇÃO E O SUCESSO ESCOLAR

FACULDADE DE CIÊNCIAS UNIVERSIDADE DE LISBOA

Inquérito

Não podemos esquecer que a proposta alimentar ao nível das cantinas e bufetes devem obedecer aos princípios de uma alimentação saudável.

Este questionário destina-se a apurar qual o tipo de alimentação que os estudantes têm actualmente, bem como os gastos envolvidos na mesma. Deste modo, pretende-se efectuar uma avaliação preliminar dos nutrientes, nomeadamente os que são essenciais para uma actividade cerebral saudável, tais como: a atenção, a memória, as funções lógicas e o controle de emoções e de comportamento.

A sua colaboração é muito importante, para que os resultados sejam credíveis.

Desde já fica o meu sincero agradecimento pela sua colaboração.

Responda às seguintes questões, colocando, se for o caso uma cruz **[X]** ou um número**,** no quadrado correspondente:

1. Data de Nascimento: ------ / ------- / ---------

2. Sexo : Feminino ☐ Masculino ☐

3. Residência:

Cidade de Lisboa ☐ Área Metropolitana de Lisboa ☐ Outra localidade ☐

3.1. Qual a duração da viagem da sua casa à faculdade_____

4. Qual o tipo de habitação em que vive?

Habitação dos pais	☐	Quarto alugado	☐	Casa alugada	☐
Residência escolar	☐	Outros	☐		

5. Indique a sua situação, quanto ao Agregado Familiar:

Vive sozinho ☐ Vive com pai / mãe ☐ Vive com outros familiares ☐

Outras situações ☐

6. Refira o seu suporte de sustentabilidade:

Bolsa atribuída ☐

Mesada fornecida pelos pais ☐

Financiado pelo próprio (trabalhador) ☐

Outras ☐

7. Quanto às suas habilitações literárias, está a frequentar:

Licenciatura ☐ Mestrado ☐ Doutoramento ☐

A ALIMENTAÇÃO E O SUCESSO ESCOLAR

8. Quais são os factores de decisão que o levam a optar pelo lugar onde toma as refeições?

Alimentação equilibrada	☐	Confiança na confecção	☐	Ementa	☐
Tempo de Espera	☐	Preço	☐	Outros factores	☐

9. Quantas refeições geralmente toma na faculdade e quantas vezes por semana?

- Duas refeições (pequeno-almoço, almoço) ☐ Número de vezes ☐

- Três refeições (pequeno-almoço; almoço; lanche) ☐ Número de vezes ☐

- Quatro refeições (pequeno-almoço; almoço, lanche, jantar) ☐ Número de vezes ☐

10. Ao pequeno-almoço, quantas vezes por semana consome estes tipos de alimentos?

Leite	☐	Manteiga	☐	Queijo	☐	Iogurtes	☐
Pão	☐	Fiambre	☐	Fruta	☐	Sumo	☐
Cereais de pequeno-almoço	☐	Bolos	☐	Outros.	☐		

11. Ao almoço, a sua alimentação é constituída por:

Sopa ☐

Prato { *Carne* ☐ / *Peixe* ☐ / *Vegetariano* ☐ }

Sobremesa ☐

Água ☐ Bebidas Alcoólicas ☐ Sumos ☐

12. Ao jantar, opta por uma refeição semelhante ao almoço?

SIM ☐ **Não** ☐

13. Quantas vezes por semana consome estes tipos de nutrientes:

Alimentos ricos em hidratos de carbono…… ☐

Alimentos ricos em ferro…….. ☐

Alimentos ricos em ácido fólico…… ☐

Alimentos ricos em vitaminas…… ☐

14. Por semana qual a quantia gasta em alimentação? ☐

A ALIMENTAÇÃO E O SUCESSO ESCOLAR

15. Qual é o ano que frequenta? ☐

16. Concluiu o ano com sucesso a todas as cadeiras? ☐

16.1. Em caso negativo, indique quantas. ☐ em ☐

17.Se é aluno de primeira matricula qual foi a sua media de acesso à faculdade ☐

FIM

Apêndice B

Inquérito
(FCUL, questões viciadas)

A ALIMENTAÇÃO E O SUCESSO ESCOLAR

Inquérito

Este questionário destina-se a apurar qual o tipo de alimentação que os estudantes têm actualmente, bem como os gastos envolvidos na mesma. Deste modo, pretende-se efectuar uma avaliação preliminar dos nutrientes, nomeadamente os que são essenciais para uma actividade cerebral saudável, tais como: a atenção, a memória, as funções lógicas e o controle de emoções e de comportamento.

A sua colaboração é muito importante, para que os resultados sejam credíveis.

Desde já fica o meu sincero agradecimento pela sua colaboração.

A sua colaboração é muito importante, para que os resultados sejam credíveis.

Desde já fica o meu sincero agradecimento pela sua colaboração.

1. Data de Nascimento ------/ -------/ ---------

Responda às seguintes questões, colocando, se for o caso uma cruz [X] ou um número, no quadrado correspondente:

1. **Sexo** : Feminino ☐ Masculino ☐

3. Residência:

Cidade de Lisboa ☐ Área Metropolitana de Lisboa ☐ Outra localidade ☐

Qual a duração da viagem da sua casa à faculdade_____

4. Qual o tipo de habitação em que vive?:

Habitação dos pais ☐ Quarto alugado ☐ Casa alugada ☐

Residência escolar ☐ Outros ☐

5. Agregado Familiar

Vive sozinho ☐ vive com pai \ mãe ☐ vives com outros familiares ☐

Outras situações ☐

6. Suporte sustentabilidade

Bolsa atribuída ☐

Mesada fornecida pelos pais ☐

Financiado pelo próprio (trabalhador) ☐

Outras ☐

7. Quanto às suas habilitações literárias, está a frequentar:

Licenciatura ☐ Mestrado ☐ Doutoramento ☐

A ALIMENTAÇÃO E O SUCESSO ESCOLAR

8. Quais são os factores de decisão que o levam a optar pelo lugar onde toma as refeições?

Alimentação equilibrada	☐	Confiança na confecção	☐	Ementa	☐
Tempo de Espera	☐	Preço	☐	Outros factores	☐

9. Quantas refeições geralmente toma na faculdade e quantas vezes por semana?

- Duas refeições (pequeno-almoço, almoço) ☐ Número de vezes ☐

- Três refeições (pequeno-almoço; almoço; lanche) ☐ Número de vezes ☐

- Quatro refeições (pequeno-almoço; almoço, lanche, jantar) ☐ Número de vezes ☐

10. É importante relembrar que, quando nos levantamos, o nível de calorias e nutrientes está muito baixo pelo que devemos tomar um pequeno – almoço equilibrado ao qual devem corresponder 25% das necessidades energéticas diárias.

Ao pequeno – almoço quantas vezes por semana consomem estes tipos de alimentos.

Leite	☐	Manteiga	☐	Queijo	☐	Iogurtes	☐
Pão	☐	Fiambre	☐	Fruta	☐	Sumo	☐
Cereais de pequeno-almoço	☐	Bolos	☐	Outros.	☐		

11.Ao almoço a sua alimentação é baseada:

Sopa ☐

Prato { *Carne* ☐ / *Peixe* ☐ / *Vegetariano* ☐ }

Sobremesa ☐

Água ☐ Bebidas Alcoólicas ☐ Sumos ☐

12.Ao jantar opta por uma refeição semelhante ao almoço

SIM ☐ **Não** ☐

13. O stress, frequente nas épocas de exames, testes, provas, pode levar a perdas de apetite, porém é necessário um esforço para que o desempenho intelectual prossiga sem sobressaltos. Nesta época mencione quantas vezes por semana consome estes tipos de nutrientes:

Alimentos ricos em hidratos de carbono ☐

Alimentos ricos em ferro…….. ☐

Alimentos ricos em ácido fólico…… ☐

Alimentos ricos em vitamina…… ☐

A ALIMENTAÇÃO E O SUCESSO ESCOLAR

14. Por semana qual a quantia gasta em alimentação ☐

15. Qual é o ano que frequenta?

16. Concluiu o ano com sucesso a todas as cadeiras? ☐

16.1. Em caso negativo, indique quantas. ☐ em ☐

17.Se é aluno de primeira matricula qual foi a sua media de acesso à faculdade ☐

Considerações finais

Muitos outros nutrimentos intervêm também no bom funcionamento regulação do sistema nervoso, reflectindo-se na concentração e na memória. É o caso do lítio, silício, selénio e crómio que são ajudantes na natureza e, como as quantidades de que necessitamos são muito pequenas, facilmente serão suprimidas desde que ingira diariamente legumes e frutas.

FIM

Apêndice C

Inquérito
(Secundário, questões neutras)

A ALIMENTAÇÃO E O SUCESSO ESCOLAR

Inquérito

Não podemos esquecer que a proposta alimentar ao nível das cantinas e bufetes devem obedecer aos princípios de uma alimentação saudável.

Este questionário destina a apurar qual o tipo de alimentação que os estudantes, bem como os gastos envolvidos, para uma avaliação preliminar dos nutrientes nomeadamente dos essenciais para uma actividade cerebral saudável (atenção, memória, funções lógicas, controle de emoções e de comportamento.

A sua colaboração é muito importante, para que os resultados sejam credíveis.

Desde já fica o meu sincero agradecimento pela sua colaboração.

1.Data de Nascimento ------/ -------/ ---------

Responda às seguintes questões, colocando uma **X** no quadrado corresponde

2.Sexo : F ☐ M ☐

3. Residência:

Cidade de Lisboa ☐ Área Metropolitana de Lisboa ☐ Outra localidade ☐

Qual a duração da viagem da sua casa à escola_____

4.Qual o tipo de habitação em que vive?

Habitação dos pais ☐ Quarto alugado ☐ Casa alugada ☐

Outros ☐

5.Agregado Familiar

Vive sozinho ☐ vive com pai \ mãe ☐ vives com outros familiares ☐

Outras situações ☐

6.Suporte sustentabilidade

Bolsa atribuída ☐

Mesada fornecida pelos pais ☐

Financiado pelo próprio (trabalhador) ☐

Outras ☐

7.Habilitações literárias.

Está a frequentar

Curso de Educação e Formação de Jovens ☐ Cursos Profissionais ☐

Aprendizagem ☐ Cursos de Educação e Formação para Adultos ☐

A ALIMENTAÇÃO E O SUCESSO ESCOLAR

8. Quais são os factores de decisão que o levam a optar pelo lugar onde toma as refeições?

Alimentação equilibrada	☐	Confiança na confecção	☐	Ementa	☐
Tempo de Espera	☐	Preço	☐	Outros factores	☐

9. Quantas refeições geralmente toma na faculdade e quantas vezes por semana?

- Duas refeições (pequeno-almoço, almoço) ☐ Número de vezes ☐

- Três refeições (pequeno-almoço; almoço; lanche) ☐ Número de vezes ☐

- Quatro refeições (pequeno-almoço; almoço, lanche, jantar) ☐ Número de vezes ☐

10. Ao pequeno – almoço quantas vezes por semana consomem estes tipos de alimentos.

Leite	☐	Manteiga	☐	Queijo	☐	Iogurtes	☐
Pão	☐	Fiambre	☐	Fruta	☐	Sumo	☐
Cereais de pequeno-almoço	☐	Bolos	☐	Outros.	☐		

11.Ao almoço a sua alimentação é baseada:

Sopa ☐

Prato { *Carne* ☐ / *Peixe* ☐ / *Vegetariano* ☐ }

Sobremesa ☐

Água ☐ Bebidas Alcoólicas ☐ Sumos ☐

12. Ao jantar opta por uma refeição semelhante ao almoço

SIM ☐ **Não** ☐

13. Quantas vezes por semana consomem estes tipos de nutrientes:

Alimentos ricos em hidratos de carbono ☐

Alimentos ricos em ferro…….. ☐

Alimentos ricos em ácido fólico…… ☐

Alimentos ricos em vitamina…… ☐

☐

A ALIMENTAÇÃO E O SUCESSO ESCOLAR

FACULDADE DE CIÊNCIAS UNIVERSIDADE DE LISBOA

14. Por semana qual a quantia gasta em alimentação

15. Avaliação

Se é aluno com a segunda matricula ou mais; indique:

- Quantos módulos efectuaram com sucesso nos anos anteriores ☐ no total ☐
- Se é aluno do primeiro ano, indique a média do curso que conclui ☐

FIM

Apêndice D

Inquérito
(Secundário, questões viciadas)

A ALIMENTAÇÃO E O SUCESSO ESCOLAR

Inquérito

Este questionário destina-se a apurar qual o tipo de alimentação que os estudantes têm actualmente, bem como os gastos envolvidos na mesma. Deste modo, pretende-se efectuar uma avaliação preliminar dos nutrientes, nomeadamente os que são essenciais para uma actividade cerebral saudável, tais como: a atenção, a memória, as funções lógicas e o controle de emoções e de comportamento.

A sua colaboração é muito importante, para que os resultados sejam credíveis.

Desde já fica o meu sincero agradecimento pela sua colaboração.

A sua colaboração é muito importante, para que os resultados sejam credíveis.

Desde já fica o meu sincero agradecimento pela sua colaboração.

1. Data de Nascimento ------/ -------/ ---------

Responda às seguintes questões, colocando, se for o caso uma cruz [X] ou um número, no quadrado correspondente:

1. **Sexo** : Feminino ☐ Masculino ☐

3. Residência:

Cidade de Lisboa ☐ Área Metropolitana de Lisboa ☐ Outra localidade ☐

Qual a duração da viagem da sua casa à escola _____

4. Qual o tipo de habitação em que vive?

Habitação dos pais ☐ Quarto alugado ☐ Casa alugada ☐

Residência escolar ☐ Outros ☐

5. Agregado Familiar

Vive sozinho ☐ vive com pai \ mãe ☐ vives com outros familiares ☐

Outras situações ☐

6. Suporte sustentabilidade

Bolsa atribuída ☐

Mesada fornecida pelos pais ☐

Financiado pelo próprio (trabalhador) ☐

Outras ☐

7. Quanto às suas habilitações literárias, está a frequentar

Cursos de Educação e Formação de Jovens ☐ Aprendizagem ☐

Cursos de Educação e Formação para Adultos ☐ Cursos Profissionais ☐

A ALIMENTAÇÃO E O SUCESSO ESCOLAR

8. Quais são os factores de decisão que o levam a optar pelo lugar onde toma as refeições?

Alimentação equilibrada	☐	Confiança na confecção	☐	Ementa	☐
Tempo de Espera	☐	Preço	☐	Outros factores	☐

9. Quantas refeições geralmente toma na faculdade e quantas vezes por semana?

- Duas refeições (pequeno-almoço, almoço) ☐ Número de vezes ☐

- Três refeições (pequeno-almoço; almoço; lanche) ☐ Número de vezes ☐

- Quatro refeições (pequeno-almoço; almoço, lanche, jantar) ☐ Número de vezes ☐

10. É importante relembrar que, quando nos levantamos, o nível de calorias e nutrientes está muito baixo pelo que devemos tomar um pequeno – almoço equilibrado ao qual devem corresponder 25% das necessidades energéticas diárias.

Ao pequeno – almoço quantas vezes por semana consomem estes tipos de alimentos.

Leite	☐	Manteiga	☐	Queijo	☐	Iogurtes	☐
Pão	☐	Fiambre	☐	Fruta	☐	Sumo	☐
Cereais de pequeno-almoço	☐	Bolos	☐	Outros.	☐		

11.Ao almoço a sua alimentação é baseada:

Sopa ☐

Prato { *Carne* ☐ / *Peixe* ☐ / *Vegetariano* ☐ }

Sobremesa ☐

Água ☐ Bebidas Alcoólicas ☐ Sumos ☐

12.Ao jantar opta por uma refeição semelhante ao almoço

SIM ☐ **Não** ☐

13. O stress, frequente nas épocas de exames, testes, provas, pode levar a perdas de apetite, porém é necessário um esforço para que o desempenho intelectual prossiga sem sobressaltos. Nesta época mencione quantas vezes por semana consome estes tipos de nutrientes:

Alimentos ricos em hidratos de carbono ☐

Alimentos ricos em ferro…….. ☐

Alimentos ricos em ácido fólico…… ☐

Alimentos ricos em vitamina…… ☐

A ALIMENTAÇÃO E O SUCESSO ESCOLAR

14. Por semana qual a quantia gasta em alimentação ☐

15. Avaliação.

Se é aluno com a segunda matricula ou mais; indique:

- **Quantos módulos efectuaram com sucesso nos anos anteriores** ☐
  no total ☐
- **Se é aluno do primeiro ano, indique a média do curso que conclui.** ☐

Considerações finais

Muitos outros nutrimentos intervêm também no bom funcionamento regulação do sistema nervoso, reflectindo-se na concentração e na memória. É o caso do lítio, silício, selénio e crómio que são ajudantes na natureza e, como as quantidades de que necessitamos são muito pequenas, facilmente serão suprimidas desde que ingira diariamente legumes e frutas.

FIM

Apêndice E

Bibliografia

Aleixo, S., Brilhante, M. F., Diamantino, F, Mendonça, S., and Pestana, D. (2007). Non-Response and Sample Size, *Bulletin of the International Statistical Institute* **LXII**, 4804–4807.

Aleixo, S., Brilhante, M. F., Diamantino, F, Mendonça, S., and Pestana, D. (2011). Erros não amostrais — uma floresta de enganos, Bol. SPE, 53–68.

Assael,M., and Keon, J. (1982). Non-sampling vs sampling errors in survey research, *J. Marketing* **46**, 114–123.

Barnett, V. (2002) *Sample Surveys: Principles and Methods*, 3rd ed., Arnold, London.

Bethlehem, J. (2011). Nonresponse in surveys, in Lovric, M. (ed.) *International Encyclopedia of Statistical Science*, Springer, New York, 982 –983.

Biemer, P. P. (2010) Overview of design issues: total survey error. In Marsden, P., and Wright, J. (eds.) *Handbook of Survey Research*, 2nd ed., Cap. 2, Bingley, United Kingdom.

Bourque, L. B., and Fielder, E. P. (2003). *How to Conduct Self-Administered and Mail Surveys*, Sage Publ., Thousand Oaks.

Bourque, L. B., and Fielder, E. P. (2003). *How to Conduct Telephone Surveys*, Sage Publ., Thousand Oaks.

Bradburn, N. M., and Sudman, S. (1979). *Improving Interview Method and Questionnaire Design*, Jossey–Bass, San Francisco.

Chandra, T. K. (1999). *A First Course in Asymptotic Theory of Statistics*, Narosa, New Delhi.

Churchill Jr., G. A. (1983). *Martketing Research — Methological Foundations*, 3rd ed., The Dryden Press.

Cochran, W. J. (1977). *Sampling Techniques*, 3rd ed., Wiley, New York.

Deming, W. E. (1944). On errors in surveys. *Am. Sociol. Rev.* **9**, 359–369.

Diamantino, M. F. (2008). *Contribuição ao Estudo de Dados em Falta*, dissertação de doutoramento, Universidade de Lisboa.

Elliot, D. (1991). *Weighting for Non-Response*, OPCS, London.

Erdös, P., and Rényi, A. (1959). On a central limit theorem for samples from a finite population, *Publ. Math. Institut. Hungar. Acad. Sci.* **4**, 49–61.

Fink, A. (2003). *The Survey Handbook*, Sage Publ., Thousand Oaks.

Fink, A. (2003). *How to Ask Survey Questions*, Sage Publ., Thousand Oaks.

Fink, A. (2003). *How to Design Survey Studies*, Sage Publ., Thousand Oaks.

Fink, A. (2003). *How to Sample in Surveys*, Sage Publ., Thousand Oaks.

Fink, A. (2003). *How to Manage, Analyze, and Interpret Survey Studies*, Sage Publ., Thousand Oaks.

Fink, A. (2003). *How to Report on Surveys*, Sage Publ., Thousand Oaks.

Hansen, M. M., and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*Ê**14**, 333–362.

Hansen, M. M., and Hurwitz, W. N. (1946). The problem of nonresponse in sample survys, *J. Am. Statist. Assoc.* **41**, 517–529.

Hansen, M. M., Hurwitz, W. N., and Madow, W. G. (1962). *Sample Survey Methods and Theory*, Wiley, New York.

Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe , *Journal of the American Statistical Association* **47**, 663–685.

Kalton, G., and Anderson, D. (1986). Sampling rare populations, *J. Royal Statist. Soc.* **A 149**, 65–82.

Kovalenko, I. N. (1965). On a class of limit distributions for rarefied flows of homogeneous events, *Lit. Mat. Sbornik* **5**, 569–573. (*Selected Transl. Math. Statist. and Prob.* **9**, Providence, Rhode Island, 1971, 75– 81.)

Kozubowski, T. J. (1994). Representation and properties of geometric stable laws, *Approximation, Probability, and Related Fields*, Plenum, New York. 321–337.

Levy, and Lemeshow, (2009). *Sampling of Populations: Methods and Applications*, 4th ed., Wiley, New York.

Lohr, S. L. (1999). *Sampling: Design and Analysis*, Duxbury, Pacific Grove.

Lovric. M. (ed.) (2011). *International Encyclopedia of Statistical Science*, Springer, New York.

Manfreda, K. L., Berzelak, N., and Vehovar, V. (2011). Nonresponse in Web surveys, in Lovric, M. (ed.) *International Encyclopedia of Statistical Science*, Springer, New York, 984–987.

Martins, J. P. (2009). *Feira dos Momentos $\tilde{N}$ Planeamento Experimental e Investigacão de Localizacão e Escala em Populacões não Gaussianas*, dissertação de doutoramento, Universidade de Lisboa.

Mosteller, F. (1978). Errors I: nonsampling errors. In: Kruskal, W. H., and Tanur, J. M. (eds.) *International Encyclopedia of Statistics*. Free Press, New York, 208–229.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. Roy. Stat. Soc.* **97**, 558–606.

Oishi, S. M. (2003). *How to Conduct In-Person Interviews for Surveys*, Sage Publ., Thousand Oaks.

Pestana, D., e Velosa, S. (2010). *Introdução à Probabilidade e à Estatística*, Fundação Gulbenkian, Lisboa.

Rényi, A. (1956). A characterization of the Poisson process, *MTA Mat. Kut. Int. Közl.* **1**, 519–527. (Reeditado em *Selected Papers of Alfréd Rényi*, P. Turán, ed, Akadémiai Kiadó, Budapest, 1976, vol. I, p. 622–628.)

Scheaffer, R. L., Mendenhall III, W., and Ott, R. L (1996) *Elementary Survey Sampling*, 5th ed., Duxbury Press, Belmont.

Singh, S. (2003). *Advanced Sampling Theory with Applications, How Michael 'Selected' Amy*, Kluwer, Dordrecht.

Tanur, J. M. (2011). Nonsampling errors in surveys, in Lovric, M. (ed.) *International Encyclopedia of Statistical Science*, Springer, New York, 988–991.

Thompson, S. K. (1992). *Sampling*, Wiley, New York.

Tryfos, P. (1996). *Sampling Methods for Applied Research — Text and Cases*, Wiley, New York.