STATISTICS DENMARK

Statistics Norway
Statistisk sentralbyrå

Statistiska centralbyrån
Statistics Sweden

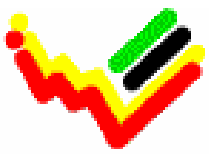
MZ:2005:2

Documentação do curso de amostragem

17-23 February 2005

TA for the Scandinavian Support Program to Strengthen the Institutional Capacity of the National Statistics, Mozambique

Kenny Petersson
Irene Tuveng



Instituto Nacional de Estatística

Ref. no. 104.Moz.98

March 2005

Irene Tuveng
Statistics Norway
Irene.Tuveng@ssb.no
+47-21 09 42 92

Kenny Petersson
Statistics Sweden
Kenny.Petersson@scb.se
+46 –19 17 65 62

Conteúdo

List of abbreviations

CO	Scanstat Coordination Office in Statistics Denmark
CAE	Classificação de Actividades Económicas (type of economic activity)
CEMPRE	Censo de empresas 2002 (Business census 2002)
CNBS	Classificação Nacional de Bens e Serviços (product classification)
Danida	Danish International Development Assistance
DEBA	Department for Statistics on Goods and Environment
DESC	Department for Statistical Services and Business Statistics
DESE	Directorate of Statistics on Enterprises and Sector Economics
DISI	Department of Informatics and Information Systems
DPINE	Provincial delegation of INE
FUE	Ficheiro de unidades estatísticas (software for the business register bought from INE-P)
INE	Instituto Nacional de Estatística, Moçambique
INE-P	Instituto Nacional de Estatística, Portugal
Scanstat	Consortium between Statistics Denmark, Statistics Norway and Statistics Sweden
SCB	Statistics Sweden
SEN	The national statistical system
SSB	Statistics Norway

1 SUMMARY IN ENGLISH

This report is a separate appendix to the report from a course in sampling that was held 17-23 February 2005 for 13 staff members of National Institute of Statistics in Mozambique (INE). The objective of producing a separate appendix was to give the participants in the course an extensive documentation, and also to facilitate the use of the same material for new courses in the same subject area. The course was held during five days in a room with 8 computers for computer training in the Portuguese school in Maputo.

The computers that were used in the course had software in Portuguese. The part of the course that refers to functions in Excel has references to the corresponding name of the function in the English language version of Excel. The part with solutions in Microsoft Access refers only to the English language version of Microsoft Access.

The course was designed to enable the participants to define a sampling frame, realize the sampling plan and process the survey in a statistically correct way. The main focus was on practical work on translating the statistical tasks to operational data processing routines. Considering the limited experience in using statistical software packages, Microsoft Excel was used for all exercises.

Using a file called FUE_MINI, with the same structure as the business register of Mozambique, the participants made a proper identification of a sample frame, calculated sample size recommendations using optimum allocation, made a sample design and executed the sampling. The exercises were made in Excel, but an Access version of FUE_Mini was also developed and used for demonstrations during the course.

The objectives of the course were to

- Give a summary of the theory for sample surveys
- Give practical skills in defining a sample frame and designing a stratification plan
- Give practical skills in calculating measures as standard deviation and confidence interval for samples and sample frames.
- Give practical skills in using optimum allocation for optimising the sample distribution between strata.
- Make samples of the most frequent types of sampling methods with main focus on stratified random samples
- Give training in how to present error related to estimates in analysis
- Emphasise the differences in the treatment of different types of missing data in sample survey and censuses.
- Give practical skills in using Excel for the realizing the most important tasks in the sampling process.

The course was held by Ms Irene Tuveng from Statistics Norway (SSB), Mr Kenny Petersson from Statistics Sweden (SCB) and Mr Firmino Guiliche from the National Institute of Statistics in Mozambique (INE).

2 RESUMO

Este relatório foi elaborado como um anexo separado do relatório sobre um curso de amostragem que foi realizado durante 5 dias em Fevereiro 2005 para um grupo de 13 funcionários do Instituto Nacional de Estatística de Moçambique (INE). O objectivo de produzir um anexo separado foi de fornecer uma documentação mais detalhada para os participantes do curso e também para dar o INE um exemplo de "manual" para facilitar as preparações para novos cursos na mesma área.

Durante o curso elaborou-se demonstrações e exercícios para elaborar amostras dos tipos mais frequentes

- amostra aleatória simples sem e com estratificação
- amostra aleatória simples sistematicamente sem e com estratificação

Aplicou-se as ferramentas para calcular medidas com média, desvio padrão e intervalo de confiança.

Os exercícios foram executados usando Excel mas demonstrou-se também como realizar o resultado correspondente em Access. Para calcular a distribuição óptima de amostras estratificadas utilizou-se a formula para "Optimum allocation" assumido que o custo por unidade da amostra é igual.

Os objectivos do curso estavam a

- dar um resumo da teoria para amostragem
- dar ê habilidades práticas em definir uma fonte de amostragem e em elaborar um plano de estratificação
- dar habilidades práticas em calcular medidas como o intervalo do desvio padrão e de intervalos de confiança para amostras e a fonte de amostragem.
- dar habilidades práticas em calcular a distribuição óptima da amostra entre estratos.
- elaborar amostras dos tipos mais frequentes de métodos de amostragem enfocando amostras aleatórias estratificadas
- dar treinamento em como tratar e apresentar os erros relacionado às estimativas na análise
- sublinhar as diferenças mais importantes no tratamento de tipos diferentes de faltas no processamento de amostras comparado de recenseamentos.
- dar habilidades práticas no uso de Excel para realizar as tarefas as mais importantes da cada etapa do processo de amostragem.

3 Teoria de estatística

A cronologia das diferentes partes do curso é conforme o ordem dos slides da apresentação de cada dia em Powerpoint.

4 Probabilidades

A parte teórica do curso dá uma introdução da teoria sobre distribuições probabilisticas e distribuições de amostras aleatórias. Uma amostra baseada duma população dá informação sobre as unidades da população, primeiro da unidade escolhida, mas também sobre os outros. Se existe uma grande quantidade de unidades de um tipo, é mais provável que este tipo também entra na amostra. Por meio de elaborar amostras aleatoriamente é possível calcular a probabilidade para cada unidade de ser escolhida e elaborar estimações sobre o total da população.

Uma observação que utiliza-se na estatística é que a média duma amostra probabilistica dá uma estimação da média da população, e quando aumenta-se a amostra aumenta-se a probabilidade que a média da amostra é perto da média da população.

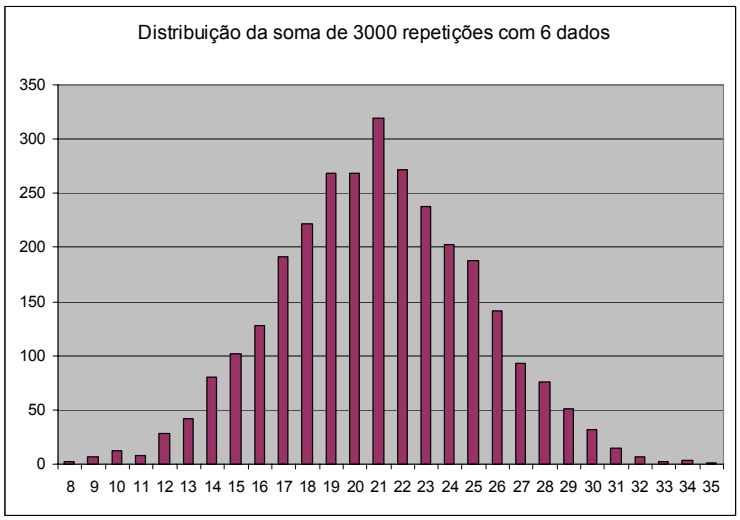
Uma explicação é que quando faz-se um grande número de amostras escolha-se cada unidade (mais ou menos) o mesmo número de vezes. Um caso especial é quando escolhemos cada unidade exactamente uma vez (o qualquer número de vezes) temos a definição da média (a soma divisão o número de observações). É fácil verificar que a média de todas amostras possíveis é igual da média da população.

No ficheiro "D21_Exercicio_TodasAmostrasPossiveis.xls" há um exemplo que faz o cálculo de todas amostras possíveis do tamanho 2 duma população de 4 unidades. Neste caso existe 6 diferentes amostras possíveis (1+2, 1+3,1+4,2+3,2+4 e 3+4).

A teorema do limite central diz que

- independentemente da distribuição de uma população aproximar ou não á distribuição normal, a distribuição amostral das suas médias fica cada vez mais próxima da distribuição normal à medida que o tamanho da amostra aumenta.

Como exemplo durante o curso mostrou-se a distribuição de 3000 repetições de jogar dados usando seis dados cada vez (ficheiro "C2_Exemplo_6Dados3000repetições.xls").



Neste caso é verdadeiro que cada valor 1-6 tem a mesma probabilidade para contribuir para o resultado final, mas uma amostra de 6 repetições frequentemente dá resultados de grande desvio do resultado médio que é 21 (=1+2+3+4+5+6).

5 A população alvo e a fonte de amostragem

Frequentemente não é possível amostrar da população que está definida como a população alvo do inquérito. A fonte de amostragem (população-grelha) pode ser uma lista de unidades que não corresponde à população alvo. No caso do FUE, o ficheiro de unidades estatísticas, o cadastro de empresas de Moçambique, existe informação actualizada durante o Censo de Empresas (CEMPRE) 2002, mas o FUE não esta actualizada com todas novas empresas depois do CEMPRE.

6 Definir a fonte de amostragem

Como definir a fonte de amostragem depende da informação disponível e o objecto do inquérito que queremos preparar. Durante o curso utilizou-se uma base de dados com a mesma estrutura como o FUE. As variáveis utilizadas para definir a fonte de amostragem eram CAE (Classificação de Actividade Económica), número de pessoal no CEMPRE (NPS), Forma Jurídica (FJR) e Situação Actual (STA).

Para realizar um plano de amostragem elaborou-se soluções baseadas de tabelas de ajuda. Quando a tarefa é de seleccionar grupos por tipo de situação actual utiliza-se uma tabela de todos os códigos de "Situação Actual" onde aplica-se as regras da amostragem por meio de indicar o grupo e/ou um indicador que indica se a secção esta incluída na fonte de amostragem ou não.

STA	STA_DSG	STA_INCL
00	SITUAÇÃO INDEFINIDA	0
01	AGUARDANDO INÍCIO DE ACTIVIDADE	0
02	EM ACTIVIDADE	1
03	PARALISADA (ACTIVIDADE SUSPENSA)	0
04	EXTINTA (CESSAÇÃO DEF. P/OUTRAS RAZÕES)	0
05	OUTRA	0

A maneira aplicar a regra em Excel é incluir uma nova coluna da base de amostragem nomeada STA_INCL e actualizar a coluna usando a formula PROCV (inglês VLOOKUP). Em Access utiliza-se uma query simples do tipo "Update query" (exemplo query "QG03_FUE_Mini_STAIncl_Actualizar" no FUE_MINI.MDB no material do curso.

7 Agrupar registos e avaliar alternativos de esquema de amostragem

A ferramenta mais útil para avaliar uma ideia de esquema de amostragem é que rever a estrutura usando uma tabela dinâmica (pivot table). Durante o curso elaborou-se grupos por CAEs e tamanho de pessoal no FUE por meio de incluir novas colunas numa tabela de empresas/estabelecimentos. No ficheiro FUE_Mini elaborou-se os estratos F1,G1,G2,H1 e H2 para grupos de CAE e os estratos "A", "B" e "C" para o tamanho do pessoal. Por meio de juntar os códigos (função CONCATENAR em Excel, Inglês CONCATENATE) elaborou-se um código de estrato para cada unidade da população.

8 Calcular a distribuição óptima do tamanho da amostra por estrato

Usando uma tabela dinâmica em Excel (ficheiro "D31_AmostraDesenhoNeyman.xls") elaborou-se cálculos da distribuição da amostra por estrato conforme "Neyman Allocation" (distribuição proporcional ao produto do número de unidade do estrato e o desvio padrão de uma variável conhecida). Os cálculos da distribuição da amostra foram elaboradas utilizando o tamanho do pessoal como indicador da homogeneidade do estrato. Quando existe velhos inquéritos na mesma área recomenda-se avaliar o desvio padrão dos indicadores centrais do velho inquérito junto com os indicadores (mais fracas) conhecidas para toda a população.

Elaborou-se também um exercício de "Neyman-Allocation" no ficheiro FUE_Mini.XLS. A folha de cálculo que conte o calculo de Neyman-allocation pode ser reutilizada como padrão para novas amostras.

9 Amostragem

A amostragem faz-se por meio de incluir um número aleatório para cada unidade da fonte de amostragem, incluir i número de unidades da amostra do estrato como uma coluna separada, ordenar a fonte por estrato e o número aleatório e depois marcar as unidades do número de ordem menor que ou igual do tamanho da amostra como amostradas. A documentação de cada etapa da amostragem de amostras aleatórias estratificadas encontra-se no ficheiro "D28_AmostraAleatóriaSimplesEstratos.xls".

Tabela 1. Exemplo amostra elaborada durante o curso

Estrato	Contar de EstID	Soma de NPS	DesvPadP de NPS
F1A	12	5 627	959,70
F1B	5	105	4,56
F1C	5	26	2,14
G1A	13	1 113	80,93
G1B	5	99	4,87
G1C	20	71	2,27
G2A	8	4 209	1248,34
G2B	5	71	3,92
G2C	40	109	2,12
H1A	7	343	13,20
H1B	5	88	3,01
H1C	5	26	2,79
H2A	2	137	18,50
H2B	5	70	4,15
H2C	20	44	1,44
Total Geral	157	12 138	419,67

10 Não resposta

O problema de não resposta é necessário tratar correctamente também quando faz-se recenseamentos totais, mas num inquérito baseado duma amostra é ainda mais importante. Por meio de assumir que as faltas tem propriedades similares que dos das respondentes pode também assumir que a média da população total é igual da média dos respondentes. Uma empresa sem actividade não dá contribuição para a produção do país mas é necessário registrar

que uma unidade da amostra não tem produção, porque dá contribuição da produção média dos respondentes. Se existe quatro unidades com produção total igual 20, a média é 5, mas se existe mais uma unidade e se sabemos que não produz nada temos informação sobre a produção para 5 unidades é a média é 20/5=4.

Normalmente não é verdadeiro que os respondentes são representativas para as unidades que não respondem. Por isso, é muito importante tentar eliminar não resposta.

Utilizou-se o ficheiro “FUE_Mini_Processamento.xls” para a discussão de não resposta.

11 Ponderadores

No ficheiro “FUE_Mini_ProcMeses.xls” mostrou-se como aplicar ponderadores. Utilizou-se também o mesmo ficheiro para discutir os efeitos do não resposta quando nenhuma unidade dum estratos responde.

Table 1. Exemplos e exercícios utilizadas durante o curso

Nome do ficheiro	Utilização/conteúdo
C2_Exemplo_6Dados3000repetições.xls	Ilustração sobre a distribuição da soma de seis dados quando faz-se 3000 repetições (exemplo de distribuição probabilistica)
D1A_Exercicio_Calculos_Simples.xls	Cálculos simples em Excel (SOMA.SE, CONTAR.SE etc..
D1C_FerramentaExcel.xls	Formulas para procurar valores em tabelas de ajuda e medidas estatísticas PROCV, INT.CONFIANCA, DESVP, etc. (em Inglês VLOOKUP, CONFIDENCE,STDDEV etc.)
D25_AmostraAleatoriaSimples.xls	Exemplos e exercícios para amostragem aleatória simples
D26_AmostraSistematicaSimples.xls	Exemplos e exercícios para amostragem aleatória simples sistematicamente
D27_AmostragemSistematicaSimplesPPS.xls	Exemplos sobre amostragem aleatória simples sistematicamente proporcional do tamanho
D28_AmostraAleatóriaSimplesEstratos.xls	Exemplos e exercícios para amostragem aleatória simples estratificada
D31_AmostraDesenhoNeyman.xls	Padrão para cálculos para optimizar a distribuição da amostra por estratos usando "Neyman-allocation"
FUE_Mini.xls	Exercício para todas etapas de preparação e elaboração de amostras aleatórias estratificadas
FUE_Mini.mdb	Padrão para executar amostras aleatórias estratificadas em Access
FUE_Mini3.xls	Segundo exercício das todas etapas de preparação e elaboração de amostras aleatórias estratificadas.
FUE_Mini_Processamento.xls	Exemplo para discussões sobre os tipos diferentes de faltas e não resposta
FUE_Mini_ProcMeses.xls	Exemplo de processamento dum inquérito baseado duma amostra e como aplicar os cálculos de ponderadores

APPENDIX 1. Programa do curso

Dato e horas	ASSUNTOS
17 de Fevereiro de 2005 08h30 - 11h00: Curso 11h00 – 11h30: Lanche 11h30 -14h00: Curso 14h00 - 14h40: Almoço 14h40: Transporte ao INE	❖ Fases duma sondagem ❖ Amostra vs. recenseamento – Vantagens / Desvantagens ❖ Introdução a amostragem probabilística ◆ Que?/Por que?/Quando?/Como? ◆ Probabilidade ◆ Amostragem Aleatória Simples (SRS) – Estimativas (médio, proporções, totais) ❖ Exemplos e exercícios em Excel: – Cálculos de probabilidades – Método de amostragem aleatória simples utilizando Excel – Cálculos de estimativas (SRS)
18 de Fevereiro de 2005 08h30 - 11h00: Curso 11h00 – 11h30: Lanche 11h30 -14h00: Curso 14h00 - 14h40: Almoço 14h40: Transporte ao INE	❖ Distribuições de variáveis aleatórios ◆ A distribuição Normal ◆ A teorema do limite central – Valor esperado E(X), Estimador não enviesado ❖ Incerteza associada com a estimativa de uma amostra: ◆ Desvio padrão ◆ Intervalo de confiança ❖ Exemplos e exercícios em Excel: ◆ Ilustração de amostragem repetida e distribuição normal dos resultados ◆ Cálculos de desvio padrão e intervalo de confiança para estimadores de SRS ❖ Amostragem aleatória sistemática ❖ Amostragem aleatória Estratificada ❖ Exemplos e exercícios em Excel: ◆ Como tirar amostras aleatória sistemática e estratificada utilizando Excel? ◆ Cálculos de desvio padrão e intervalo de confiança para estimadores de diferentes métodos de amostragem

21 de Fevereiro de 2005 08h30 - 11h00: Curso 11h00 – 11h30: Lanche 11h30 -14h00: Curso 14h00 - 14h40: Almoço 14h40: Transporte ao INE	❖ Probabilidade proporcional ao tamanho (pps) ❖ Amostragem multi-etapas ❖ Amostragem por "*clusters" (conglomerados)* ❖ Exemplos e exercícios em Excel: ◆ Repetição dos métodos de tirar amostras e cálculos de estimativas e medidas de erro utilizando um "mini-FUE" ◆ Preparação da população-grelha para amostragem (fazer códigos por estratos etc.) ❖ População-alvo - População-grelha (sample frame) ◆ Erro de cobertura ◆ Preparação da população-grelha ❖ Determinação do tamanho da amostra ❖ Distribuição da amostra por estratos ❖ Exemplos e exercícios em Excel: ◆ "Neyman-allocation" por estratos utilizando número de pessoal
22 de Fevereiro de 2005 08h30 - 11h00: Curso 11h00 – 11h30: Lanche 11h30 -14h00: Curso 14h00 - 14h40: Almoço 14h40: Transporte ao INE	❖ Exemplos e exercícios em Excel cont. ◆ "Neyman-allocation" por estratos utilizando número de pessoal. (Cont.) ◆ Demonstração em Access ❖ Erros duma sondagem: ◆ Erros de amostragem ◆ Erros não-amostrais ❖ Ponderadores e estimação
23 de Fevereiro de 2005 08h30 - 12h00: Curso 12h30 – 11h30: Almoço na "Costa de Sol"	❖ Resumo do curso ❖ Teste e avaliação do curso

APPENDIX 2. Apresentação da teoria de amostragem em PowerPoint

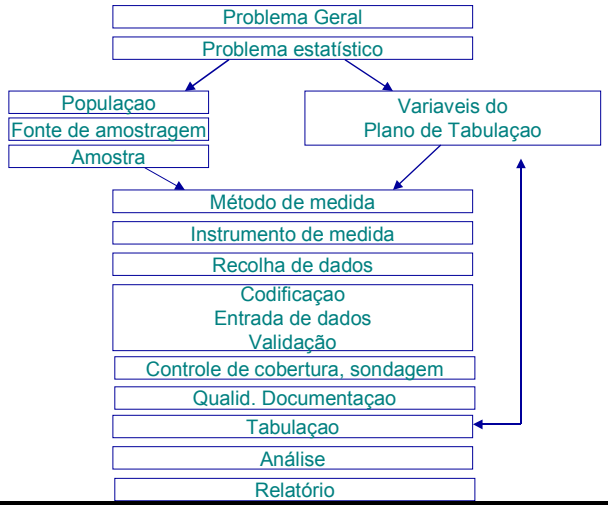
Curso de amostragem no INE

Período: 17 a 23 de Fevereiro 2005

Responsáveis pelo course:

Kenny Petersson – Estatística da Suécia (SCB) – kenny.petersson@scb.se
Irene Tuveng – Estatística da Noruega (SSB) – irene.tuveng@ssb.no

Fases de uma Sondagem



Introdução: Amostra vs recenseamento

Alguns conceitos fundamentais:

- População: conjunto finito ou infinito de elementos em relação aos quais se pretende estudar determinadas característica;
- Amostra: subconjunto da população a partir da qual se pretende tirar conclusões sobre as características da população donde provém.

Introução: Amostra vs recenseamento

Como obter a informação:

- Via administrativa;
- Recenseamento ou inquérito exaustivo, observando todos os elementos da população (muitas vezes serve de base para a extracção de amostras);
- Amostragem, observando apenas alguns elementos da população.

Introdução: Amostra vs Recenseamento

Amostra	**Recenseamento**
Menos custos	Mais custos
Rapidez	Requer mais tempo
Mais preciso	Menos preciso

Introdução: Amostra vs Recenseamento

Amostra	**Recenseamento**
Menos inquiridores melhor treinados	Mais inquiridores
Resultados menos desagregados	Resultados a um nível muito desagregado

INTRODUÇÃO A AMOSTRAGEM PROBABILÍSTICA

- QUE?
- POR QUE?
- QUANDO?
- COMO?

QUE é amostragem probabilística ?

Definição:

(i) uma metodologia de amostragem pela qual cada unidade (por exemplo 'pessoa', 'agregado familiar' ou 'empresa/estabelecimento') duma população-alvo definida tem uma possibilidade conhecida e non-zero de ser seleccionada

(ii) Procedimento em que todos os elementos da População têm uma probabilidade conhecida e superior a zero de integrar a Amostra

QUE é amostragem probabilística ?

- A chance - ou a probabilidade - de seleccionar uma unidade para inclusão na pesquisa/inquérito deve ser matematicamente calculável.
- Devem ser aplicadas técnicas probabilísticas em toda etapa do procedimento de selecção da amostra.
- Amostragem probabilística **não** é sinónimo de amostragem aleatória.

PROBABILIDADE

- Definição genérica:
  - *Verosimilhança que um acontecimento (A)aconteceu ou acontecerá.*
- Conceito matemático:
  - ***P (.)** é uma função que associa a cada acontecimento A um número no intervalo [0,1] .*

Probabilidades são figuradas de vários modos:

- *a priori* (lançamento duma moeda será 'face', P = ½ ou 0.5)
- a fracção entre o número dos resultados desejados e os resultados possíveis (e. g. dois lançamentos sucessivos serão 'face', P = ¼ ou 0.25)
- uma frequência histórica observada: (e.g. dias por ano com chuva em Bergen, Noruega, P = 275/365 ou 0.75)

Exemplo de selecção e cálculo prob. em inquéritos

- seleccionar uma amostra aleatória de 30 crianças duma lista de 1215 crianças. A probabilidade de seleccionar uma criança particular é:

$$P = \frac{n}{N} = \frac{30}{1215} = \frac{1}{40.5}$$

Exemplo de método de selecção não probabilístico

- Seleccionar uma amostra probabilística de 25 crianças da escola A para descobrir a proporção que fez as lições de casa (TPC). Você decide seleccionar as primeiras 25 crianças a chegarem à escola na terça-feira.
- *Problema*: Você não tem nenhum modo de saber sem dúvida que é igualmente provável que todas cheguem na hora.

Exemplo de método de selecção não probabilístico

- Então, as que não apareceram por motivo de transporte, ou as que não apareceram no mesmo dia por motivos de doença e as que atrasaram por outras razões têm a possibilidade zero de ser seleccionada para a amostra.
- *Por que não é isto uma amostra probabilística?*
  - A amostra representa apenas as crianças que compareceram na escola nesse dia e as que chegam à hora. Isto não é a população-alvo pretendida, porque você quer representar todas as crianças na escola, incluindo as que naquele momento estiveram ausentes.

POR QUE é a amostragem probabilística importante?

- Permite inferir os resultados à população, isto é, as conclusões podem ser generalizadas para a população inteira (universo)
- Permite que o cálculo de medidas de fiabilidade (isto é, erros de amostragem) seja feito para os resultados
- Permite apresentar os resultados numa base científica fundamentada

QUANDO deve **este método ser usado?**

A população-alvo deve ser definida sem ambiguidade com base em princípios rigorosos

ALGUNS PRINCÍPIOS **necessários :**

- Os objectivos do inquérito precisam de serem especificados claramente
- Conteúdo do inquérito: Que quer investigar?
- Variáveis analíticos: Quais estimativas quer fazer?
- Nível de desagregação: Representatividade dos dados: nacional, urbano - rural, provincial, etc.?

ALGUNS PRINCÍPIOS **necessários cont...:**

- Deve haver um *universo* (mais sobre isto mais tarde)
- A precisão requerida deve ser indicada de modo que o tamanho de amostra possa ser determinado (mais sobre isto mais tarde)
- Orçamento e constrangimentos do campo têm que ser levados em conta

COMO fazer amostragem probabilística?

- ***Métodos* de amostragem probabilística:**
- Amostragem Aleatória Simples
- Amostragem Aleatória Sistemática
- Amostragem Aleatória Estratificada
- Selecção com probabilidade proporcional ao tamanho (PPS)
- Amostragem Aleatória por Clusters (conglomerados)
- Amostragem Multi-Etapa

Amostragem Aleatória Simples - Simple Random Sampling (SRS)

- O método de amostragem mais elementar, mas usado raramente em inquéritos.
- Método da selecção é por números aleatórios, com ou sem reposição
- A maioria de teoria da amostragem é baseada no método SRS

Vantagens e Desvantagens com SRS

- *Vantagens:*
  - O mais fiável de todos os métodos de amostragem
  - Os procedimentos de selecção são simples
- *Desvantagens*
  - O método mais caro

SRS - Matemática

- População = N elementos (unidades)
- Tamanho da amostra = n
- Método de selecção com probabilidades iguais
- Probabilidade = n/N

Estimativas de SRS

Média: $\bar{x} = \sum_{i=1}^{n} x_i / n$

Proporção : $\hat{p} = \sum_{i=1}^{n} x_i / n$

Total: $x' = N \cdot \sum_{i=1}^{n} x_i / n$

Amostra de 10 estudantes;

Calcule a idade média:

x1 = 25	x2 = 31	x3 = 28
x4 = 37	x5 = 32	x6 = 40
x7 = 22	x8 = 42	x9 = 26
x10= 30		

$$\bar{x} = \frac{25 + 31 + 28 + ... + 26 + 30}{10} = \frac{313}{10} = 31.3$$

Calcule a proporção casada:

x1 = 0; x2 = 1; x3 = 1
x4 = 1; x5 = 1; x6 = 1
x7 = 0; x8 = 1; x9 = 0
x10 = 0

$$x_p = \frac{0+1+1+1+1+1+0+1+0+0}{10} = \frac{6}{10} = 0.6$$

Suponha que a população total dos estudantes, de qual a amostra de 10 foi seleccionada corresponde 75 estudantes. Então, N = 75.

Estime o número total de estudantes casados na população:

x1 = 0 x2 = 1 x3 = 1
x4 = 1 x5 = 1 x6 = 1
x7 = 0 x8 = 1 x9 = 0 x10 = 0

$$x' = \frac{75}{10}(0+1+1+1+1+1+0+1+0+0) = 45$$

A distribuição Normal – A curva "Bell"

- A distribuição normal com uma média egual a μ e um **desvio padrão** tem a notação = N(y,ó)
- A forma da curva depende o valor de desvio padrão

- SRS de n estudantes dá uma estimativa para o valor real na população, não é uma número exacto
- Assim, há incerteza associada com a estimativa de uma amostra particular.
- Para descrever esta incerteza é usando o **desvio padrão** (sd) ou **Intervalo de confiança** (I.C.)

A teorema do limite central

- Independentemente da distribuição de uma população aproximar ou não á distribuição normal, a distribuição amostral das suas médias fica cada vez mais próxima da distribuição normal à medida que o tamanho da amostra aumenta.

- O valor médio para X calculado sobre todas amostras possíveis é chamado o *valor esperado* E(X)
- Na prática, somente uma amostra única é seleccionada.
- Se esta amostra única seja uma amostra probabilística não enviesada, o valor esperado da estimativa para X = a média sobre todas amostras possíveis

A teorema do limite central

- Se **n** for suficientemente grande, o estimador **y** ~ N (Y, σ), sem pensar na distribuição subjacente das observações. Então, **y** será um estimador não enviesado de Y:
  - um *intervalo de confiança* **y** +/- 1 erro (desvio) padrão conterá Y com probabilidade ? 68 por cento
  - um *intervalo de confiança* **y** +/- 2 (1.96) erros (desvios) padrão conterá Y com probabilidade ? 95 por cento
  - um *intervalo de confiança* **y** +/- 3 erros (desvios) padrão conterá Y com probabilidade ? 99 por cento

A teorema do limite central

- O teorema de limite central é a base científica na qual o método da amostragem é baseado.
- É consequentemente a base que permite que nós tiremos conclusões a respeito da população inteira quando só uma amostra válida desta é seleccionada e investigada.
- Amostra válida = amostra probabilística

Medida do erro (desvio)

- O erro padrão duma amostra = a raiz quadrada da variância.
- Por que é o erro padrão importante?
  - Com uma estimativa do **p** + o erro padrão dela, pode-se construir um *intervalo de confiança* em torno da estimativa.
- Um intervalo de confiança permite medir a precisão de um estimador

Erro padrão - Matemática

Desvio padrão de estimativa de média

$$Sd(\bar{X}) = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

Desvio de estimativa de proporção

$$Sd(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Desvio de estimativa de total

$$Sd(X') = N \cdot \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

Erro padrão e Intervalo de confiança

- A estimativa mais/menos um erro padrão = Um intervalo de confiança de 68 por cento.
- A estimativa mais/menos dois erros padrões = Um intervalo de confiança de 95 por cento. (Exactamente 1.96 e.p)

95% I.C. – Matemática (SRS)

Média:

$$\bar{X} \pm 1{,}96\sqrt{\frac{sd(\bar{X})^2}{n}(1-\frac{n}{N})}$$

Total:

$$N\bar{X} \pm N.1{,}96\sqrt{\frac{sd(\bar{X})^2}{n}(1-\frac{n}{N})}$$

Proporção:

$$\hat{p} \pm 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}(1-\frac{n}{N})}$$

Ilustração:

- Suponha uma estimativa da proporção das pessoas empregadas é 0.6, e o calculo do erro padrão dá o resultado 0.025.
- O intervalo de confiança (nível de confiança 95 por cento) então é 0.6 ± (1.96* 0.025), ou 0.6 ± 0.05, ou [0.55, 0.65].
- Consequentemente, as possibilidades são 95 de 100 que a proporção da população (o valor verdadeiro) está no intervalo [0.55, 0.65].

Ilustração:

Erro padrão e limites de confiança da despesa, por áreas de residência, Moçambique, 2002-03

Área de Residência e Províncias	Estimativa da Despesa Diária	Desvio Padrão	Limite de Confiança		Núm. Agregados
			Inferior	Superior	
Média Nacional	**51.280**	**4,4**	**46.806**	**55.754**	**8.700**
Urbano	88.708	9,8	71.362	106.053	4.008
Rural	35.297	2,7	33.411	37.182	4.692

Os resultados do inquérito mostram que a despesa média diária por agregado familiar é de 51.280 Meticais, e o desvio padrão é de 4,4 Milhões de Meticais para um intervalo de confiança que varia entre 46.806 e 55.754 Meticais. O intervalo de confiança, indica o intervalo que pode assumir a despesa média caso o inquérito fosse repetido muitas vezes com um grau de confiança de 95%; neste caso, por cada 100 observações, 95 ficarão entre 46.806 e 55.754 Meticais.

AMOSTRAGEM ALEATÓRIA SISTEMÁTICA

- Mais ou menos equivalente a SRS (na teoria matemática)
- Método é sem reposição.
- Precisa-se de atribuir um *número de identificação* (ID) a cada unidade/elemento que compõem a população (população grelha/acessível/do estudo)

AMOSTRAGEM ALEATÓRIA SISTEMÁTICA

- Determine o *intervalo de amostragem* (IA) : A razão N/n (arredondada para o inteiro imediatamente inferior)
- Gere o *início aleatório* (R) - o número de ordem da primeira unidade a incluir na amostra.
- Integre cada IA-éssimo elemento
- O universo deve ser ordenado aleatoriamente

Vantagens e desvantagem

- *Vantagens*
  - Implementação dos procedimentos de selecção é mais simples do que SRS
  - Pode ser usado em cada etapa de amostragem
- *Desvantagem*
  - Se o universo for cíclico, os resultados podem ser enviesados

AMOSTRAGEM ALEATÓRIA ESTRATIFICADA

Metodologia:

- Informação auxiliar é usada para melhorar amostragem pela construção dos *estratos*. Um estrato é um segmento da população caracterizado por um ou mais atributos.

Variáveis de estratificação - exemplos:

- Características sócio-demográficas dos indivíduos: Sexo, idade, nível de escolaridade, classe social, classes salariais
- Características geográficas: Urbano - rural, província, região, distrito
- Características das empresas: Actividade económica da CAE, tamanho da empresa

Variáveis de estratificação

- Em geral: Variáveis correlacionadas com assuntos do inquérito
- Também: As variáveis de estratificação devem ser de boa qualidade.

Princípios da amostragem estratificada

- O número e o tipo de estratos depende a função do tamanho e da estrutura administrativa da população
- Os estratos devem ser internamente homogéneos
- Devem ser externamente heterogéneos

Vantagens e desvantagens

- *Vantagens*
  - Confiabilidade melhorada
  - Pouco efeito no nível dos custos (não aumentados)
  - Estratificação implícita é fácil implementar
  - Permite obter resultados mais eficientes com uma amostra de menor dimensão e igual representatividade (menor custo, menor tempo e menor possibilidade de erro).
- *Desvantagem*
  - Estratificação muito complexa pode ser complicado construir, com pouca melhoria na confiabilidade

MATEMÁTICA DA AMOSTRAGEM ESTRATIFICADA

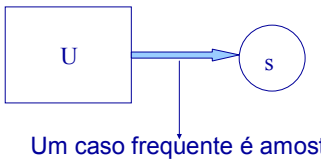
Estimativa da média ou proporção: $\bar{x} = \sum W_h \overline{x_h}$

onde $W_h = N_h/N$, a proporção da população no h-éssima estrato.

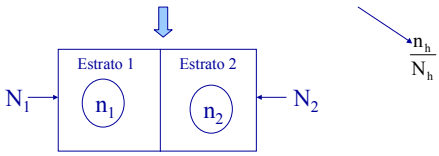
Variância da estimativa: $\sigma^2(\bar{x}) = \sum W_h^2 (1 - n_h/N_h) \frac{s_h^2}{n_h}$

onde s_h^2 é o variância no h-éssima estrato :

$$s_h^2 = \sum_{i=1}^{n_h} (x_{ih} - \overline{x_h})/(n_h - 1)$$



Um caso freqüente é amostragem aleatório por estrato
Cada unidade em cada estrato da fonte de amostragem tem probabilidade igual para ser incluído na amostra



Probabilidade proporcional ao tamanho (pps)

- Cada unidade de população-grelha tem uma probabilidade separada de ser selecionada depende um medida de tamanho das unidades
- Por exemplo podemos utilizar o numero de pessoal por empresa para desidir a probabilidade de selecção
- Amostragem aleatório estratificada pelo tamanho é um método alternativa para introduzir probabilidade depende ao tamnho

Amostragem multi-etapas - Multi-stage sampling

- Selecção duma amostra, usando mais que uma etapa
  - Primeira etapa: Selecçionar unidades primárias de amostragem. Pode ser áreas geográficos.
  - Segunda etapa: Listar todos unidades nas áreas selecçionada na primeira etapa, e tira uma amostra utilisando SRS ou PPS

Metodo**AMOSTRAGEM ALEATÓRIA POR *CLUSTERS* (Conglomerados)**

- A abordagem é seleccionar um *cluster* das empresas, geralmente definido por área geográfica
- Todas empresas no *cluster* são seleccionados
- Unidades dentro o cluster deve ser heterogéneous e representativa para a população total
- Mas geograficamente perto de um ao outro

Vantagens e desvantagem

- A razão do uso da amostragem por *clusters* é reduzir as viagens inquirindo muitas unidades dentro da mesma área geográfica, assim reduzir custos
- Frequentemente é usado em combinação com amostragem multi-etapas
- *Desvantagem*
  - Aumenta os erros de amostragem

População – Fonte de amostragem

- POPULAÇÃO – todas unidades sobre quais queremos informação. Por exemplo todas as empresas activas na área de construção. *Também designada por POPULAÇÃO - ALVO*
- FONTE DE AMOSTRAGEM- uma listagem ou outro tipo de ficheiro elaborado ou disponível para descrever a população. *Também designada por POPULAÇÃO – GRELHA (Sample frame).*

Há dois tipos de "Sample frame"

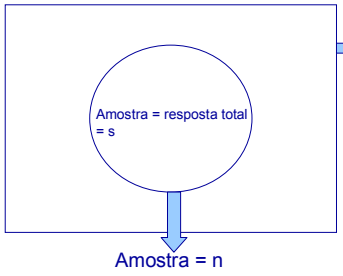
- *List frame'* é uma lista do universo, ou da população.
  - Exemplos: Uma lista de áreas administrativas; uma lista de agregados familiares numa aldeia, uma lista de empresas
- *'Area frame'* é também uma lista, mas uma lista de segmentos geográficos.
  - É construído começando com uma área geográfica grande a ser dividida em áreas pequenas, mutuamente exclusivas e exaustivas.
  - Exemplo: Áreas de enumeração do Censo (secções censitárias) são sub-áreas geográficos de um país inteiro.

Características da população-grelha

- Idealmente, a grelha deve ser:
  - Completo
  - Exacto
  - Actualizado
  - Exaustivo, sem lacunas e sem sobreposição
- Para grelhas des áreas:
  - Áreas claramente delimitadas
  - Disponível em mapas
  - Deve ter medidas do tamanho

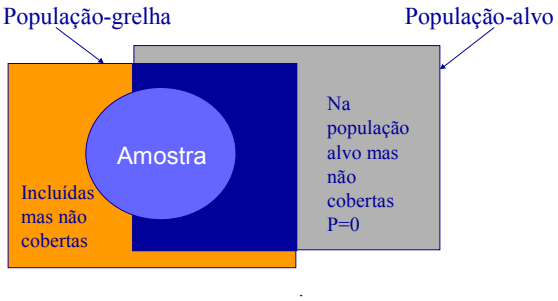
Amostragem situação ideal

População-alvo = População-grelha



O critério principal duma amostra representativa é que todas unidades da população tem probabilidade de ser incluído na amostra

Diferença entre a população alvo e a fonte de amostragem



Casos de erros da cobertura

- A fonte da amostragem não actualizada (novas empresas e empresas fechadas/extintas)
- Erros na fonte de amostragem (outro tipo de actividade principal ou número do pessoal não conhecido).
- Não existe uma fonte de amostragem que cobre toda a população alvo.

Problema de cobertura

- Estimação do nível global demais baixo
- Unidades cobertas provavelmente não são representativas para as unidades não cobertas, por exemplo novas empresas

Problemas de cobertura da população-grelha implica erros nas estimativas dos níveis e também da estrutura

TAMANHO DE AMOSTRA

- É influenciado através de:
  - Margem de erro pretendida / Nível da confiança pretendido
  - Recursos disponiveis
- Margen do erro não sabemos antes tiramos um amostra. Por isso, pretende se utilizar;
  - resultados des anos passados
  - adivinhando qualificado

Determinação do tamanho de amostra

- O tamanho de amostra requerido é diferente para cada indicador ou variável no inquérito
- Deve-se escolher um indicador *chave*, desde que somente *um* tamanho de amostra pode ser usado

Recomendações para escolher o indicador chave:

- Escolha entre dos indicadores principais do interesse no país
- Escolha esse que renderá o tamanho de amostra o maior (geralmente o indicador mais 'raro')

Tamanho da amostra - matemática

Amostragem por SRS dà um margen do erro (intervalo de confiança) dum proporção por N grande egual:

$$d = 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow n = \frac{\hat{p}(1-\hat{p})\cdot 1{,}96^2}{d^2}$$

Para proporções nos utilizarmos o facto que p(1-p) é perto de 0,25 por muitas valores de p. Por isso;

$$d = 0{,}03 \Rightarrow n = \frac{0{,}25\cdot 1{,}96^2}{0{,}03^2} = 1067$$

Tamanho da amostra -matemática

Se a sua estimativa vai ser um média um intervalo de confiaça (por N grande) :

$$d = 1{,}96\sqrt{\frac{sd(\overline{X})^2}{n}}$$

Use o mesmo procedimento, mas dê um valor a s

- adivinhe ou olhe para resultados de ano passado

Métodos de distribuir as unidades da amostra por estratos

- **Distribuição proporcional:** $n_h = \left[\frac{N_h}{N}\right] \cdot n$

- Amostragem proporcional dentro dos estratos é usado quando o objectivo do inquérito é estimativas ao nível nacional

- Amostragem desproporcional é usado quando os subgrupos têm a prioridade, isto é, o objectivo é obter estimativas para subgrupos com confiabilidade igual

Métodos de distribuir as unidades da amostra por estratos

- **Distribuição óptima:** $n_h = n \cdot \frac{N_h \sigma_h}{\sum_{h=1}^{H} N_H \sigma_h}$

Este método requer que o tamanho de amostra em cada estrato seja proporcional ao produto do tamanho de estrato e o desvio padrão de variável da nossa interese dentro o estrato.

Este método vai minimar a variança da nossa estimativa, mas ó deve ser estimada porque não é conhecido

Erros de um sondagem/inquérito

ERRO DE AMOSTRAGEM
+ ERROS NÃO-AMOSTRAIS

*= **ERRO TOTAL***

Erros "não amostrais" - poderia ser causados por:

- Erros de cobertura
- Erros de medida
- Erros de processamento
- Erro de não-resposta

Erros de medida:

- Erros do respondente (unidade na amostra):
  - mal-entendidos
  - respostas faltando à verdade
  - erros da memória
- Erros do inquiridor
  - Formulação incorrecta das perguntas
  - marcar ou anotar respostas incorrectamente
  - inventar respostas

Erros de processamento:

- Erros de codificação
- Erros de digitação
- Erros de programação
- Erros de aplicação de fórmulas de estimação
- Mal-entendidos na análise

Como minimizar erros "não amostrais"?

- Elaboração e desenho do questionário
- A necessidade do teste – piloto (inquérito piloto)
- Usar métodos de amostragem probabilística em cada etapa da selecção
- Actualize correctamente 'frames' velhos
- Treinamento intensivo dos inquiridores
- Procedimentos para assegurar a qualidade, controlar o processamento dos dados

Efeitos de não-resposta

- Major erro de amostragem
- Custos para tentar diminuir as faltas
- O problema principal - Resultados não representativas

Taxa de resposta. Analise por variáveis conhecidos

Tipo de dono	Publico	Privado
Taxa de resposta (%)	80	50

Província	1	2	3
Taxa de resposta (%)	70	80	50

Métodos para tratar não resposta

- Recalcular ponderadores no nível global
- Recalcular ponderadores por estrato
- Recalcular no base de informação suplementar
- Imputação

Ponderação e estimação

- Ponderação e também chamada 'inflação' ou 'amplificação' dos dados originais (não processados)
- Ponderação é usada na preparação de estimativas
- Ponderação tem que ser usada se a amostra não for auto-ponderada ('self-weighting')

Auto-ponderação

Definição:

- *Uma amostra é auto-ponderada quando cada unidade (pessoa/agregado familiar/empresa) é seleccionada com probabilidade igual.*

- Se a amostra for auto-ponderada, a ponderação pode ou não poderia ser usada
  - não é necessária para proporções, médias, rácios, (mas poderia ser usada)
  - é necessária para a estimação de totais

Varios tipos de ponderadores

- Ponderadores de desenho
- Ponderadores para corrigir/ajustar não-respostas
- Ponderadores para pôr as estimativas do inquérito de acordo com dados independentes

Poderadores para corrigir/ajustar não-resposta

- Ponderador a ser usado além do ponderador do desenho
- Tais ponderadores são usados para estimar totais e/ ou quando as taxas de não-resposta são diferentes por área, estrato, domínio ou cluster
- Ponderadores de desenho e ponderadores para ajustamento não-resposta são multiplicativos

Imputações

- Implica que faz-se estimação (subjectivamente) dum campo de falta no questionário
- É importante que existe códigos para indicar imputações
- O resultado é subestimação do erro de amostragem