1 População

1.1 Definição

Em sentido estatístico, pode definir-se população como um conjunto de indivíduos do mesmo tipo que diferem no que respeita a uma característica designada por atributo ou variável. Podemos estudar numa mesma população (p.e. de árvores) vários atributos (diâmetro, idade, altura). Neste caso, do ponto de vista estatístico, a população é encarada como várias populações (no exemplo, três).

Os atributos podem ser quantitativos – se podem ser directa ou indirectamente medidos – ou qualitativos – se apenas podem ser descritos e contados. Note-se que é possível converter uma característica quantitativa noutra qualitativa e vice-versa. Por exemplo, se colhermos dados de acordo com o atributo "ocorrência de árvores de uma determinada espécie com um *d* acima dos 50 cm" o atributo quantitativo "*d*" foi convertido para um atributo qualitativo. Por outro lado, atributos qualitativos como a qualidade da madeira, p.e., podem ser expressos em classes de diferente valor económico expressas em termos quantitativos.

Uma amostra é um subconjunto dos indivíduos da população que é analisado com o objectivo de conhecer as características da população. Um dos primeiros problemas que se nos põe ao planear uma amostragem é o da definição da população, ou seja, o de definir os indivíduos que a constituem e dos atributos a avaliar. Devem seleccionar-se como indivíduos os mesmos elementos que vão ser objecto de selecção para efeito de amostragem, os quais podem não coincidir com os elementos constitutivos da população em linguagem corrente. Este problema, da definição dos indivíduos que melhor se adaptam aos objectivos da amostragem, é particularmente importante nas amostragens realizadas para a avaliação de variáveis por unidade de área, bastante usual no âmbito de um inventário florestal. Tentemos aclarar este conceito com base um alguns exemplos.

1.2 Exemplos

Os exemplos apresentados ilustram bem a diversidade de problemas de amostragem na área da avaliação de recursos florestais. Note-se que alguns dos exemplos apresentados correspondem a problemas de amostragem complexos, estando a solução mais adequada a alguns deles fora do âmbito deste curso. Ao longo do texto, estas populações serão utilizadas, sempre que seja adequado, como ajuda na explicação dos diversos temas tratados.

Exemplo 1 – Avaliação do volume total num talhão de uma Mata Nacional estratificada espacialmente (população I)

(adaptado de Loetsch *et al.*, 1977)

Considere-se a figura 1, que representa um talhão de uma Mata Nacional (população I), com uma área de 40 ha. Trata-se da avaliação de um atributo que é claramente quantitativo. Para efeitos da avaliação do volume total do talhão, podemos optar por definir cada uma das árvores como indivíduo. Neste caso, a dimensão da população será igual ao número total de árvores do talhão. Não é fácil encontrar um método para seleccionar uma amostra de árvores, nem para as localizar posteriormente no campo (veremos mais à frente que há alguns esquemas de amostragem que procuram contornar este problema). Em alternativa pode dividir-se a população em pequenas parcelas, com área por exemplo de 1000 m^2, sendo cada uma destas parcelas um indivíduo da população. Neste caso, a dimensão da população será igual ao número de parcelas que couberem nos 40 ha. Na figura 1, o talhão em questão encontra-se já dividido em parcelas com uma área igual a 1000 m^2, sendo portanto a dimensão da população N=400.

O número que se encontra dentro de cada parcela corresponde ao volume da parcela expresso em m^3ha^{-1}. Neste exemplo, conhecemos o valor do atributo para a totalidade dos 400 indivíduos que constituem a população, facto que nunca acontece nas populações que amostramos na prática (senão nem sequer teríamos que amostrá-las). Para efeito de aprendizagem das técnicas de amostragem é, contudo, bastante útil trabalhar com populações em que se conhecem todos os indivíduos. É assim possível conhecer os verdadeiros valores dos parâmetros da população e compará-los, posteriormente, com as respectivas estimativas obtidas por amostragem.

Uma outra característica desta população pouco comum na prática é o facto de os limites da população coindirem exactamente com os lados das diversas parcelas que constituem os indivíduos.

	I					II					III					IV					
20	130	153	153	112	200	106	100	147	118	165	--	--	12	--	35	--	18	--	--	24	
19	124	106	136	130	165	141	194	212	136	88	100	--	12	65	88	--	100	30	12	47	
18	177	165	136	124	171	106	82	177	147	165	118	82	47	6	88	12	30	--	--	24	**A**
17	165	112	124	118	153	118	224	136	118	159	141	65	35	24	--	30	30	53	53	30	
16	100	82	118	153	147	130	130	112	88	118	147	153	88	53	71	--	--	94	47	30	
15	224	247	217	230	130	259	277	100	147	171	200	171	118	141	82	59	71	6	--	--	
14	253	200	135	271	277	271	230	206	242	177	141	200	135	153	106	153	124	71	30	6	
13	212	277	265	212	206	171	289	259	183	247	194	277	183	165	88	106	118	136	53	71	**B**
12	224	283	247	300	100	318	277	306	177	200	177	271	141	71	124	71	188	171	159	94	
11	100	141	265	277	306	165	253	265	271	159	236	188	300	165	147	241	118	159	82	124	
10	277	330	253	218	177	353	330	253	171	194	241	177	177	118	88	106	118	188	77	165	
9	224	212	159	224	141	183	283	188	147	183	206	183	130	88	59	130	141	112	106	94	
8	271	318	200	271	218	253	260	200	147	259	253	77	165	242	153	194	106	224	59	141	**C**
7	277	277	206	236	230	230	294	165	294	212	259	159	94	124	212	100	159	124	218	200	
6	130	218	65	171	165	194	171	206	312	94	153	118	171	71	136	147	88	100	153	124	
5	218	130	118	130	82	171	147	124	177	183	159	94	124	212	100	159	124	100	82	71	
4	106	147	153	118	159	153	153	130	112	177	88	12	41	18	24	88	53	41	--	18	
3	130	200	194	100	141	165	153	147	177	194	106	35	--	18	--	--	35	30	41	35	**D**
2	77	165	159	159	183	118	124	124	94	159	71	--	100	18	6	6	--	--	--	30	
1	188	183	177	130	94	153	47	188	112	118	18	18	--	--	--	12	--	30	59	12	
	a	b	c	d	e	f	g	h	I	j	k	l	m	n	o	p	q	r	s	t	

Figura 1. População I – talhão de uma Mata Nacional - constituída por 400 parcelas de 0.1 ha cada, indicando-se em cada parcela o volume em m^3ha^{-1}.

A mata a que corresponde a população I encontra-se dividida em 16 sub-talhões. Podem calcular-se as seguintes médias por talhão:

Coordenadas	I	II	III	IV
1-5	138	136	57	27
6-10	224	225	167	96
11-15	219	223	154	135
16-20	146	144	80	41

As médias dos talhões diferem consideravelmente, correspondendo esta estrutura a uma mata constituída por povoamentos de diferentes idades.

Uma análise cuidada da figura 1 permite detectar que esta população é espacialmente estratificada, ou seja, apresenta zonas bem definidas com diferentes grandezas do atributo em análise. O conhecimento do facto das populações florestais serem espacialmente estratificadas e de, portanto, as unidades individuais raramente se distribuírem ao acaso na zona a amostrar, é de importância fundamental para o planeamento e execução dos inventários florestais. Um processo de contornar a não aleatoriedade das populações florestais, tirando partido da sua estratificação espacial, é encontrar um esquema de amostragem adequado, como veremos na discussão dos delineamentos da amostragem que a seguir se discutirão.

Exemplo 2 – Avaliação do volume total num talhão de uma Mata Nacional não estratificada espacialmente (população II)

(adaptado de Loetsch *et al.*, 1977)

A figura 2 representa a população II, a qual resulta da troca da ordem das colunas na população I: a última coluna *t* segue agora a *a*, a coluna *s* segue a *b*, etc (da sequência de colunas *a b c d*... passou-se a *a t b s c*...). O objectivo desta manipulação foi simular uma população com uma distribuição mais homogénea dos volumes, sem estratificação espacial portanto.

É importante salientar que uma população muito variável não é forçosamente uma população heterógenea. Do ponto de vista da amostragem a população II é mais homogénea do que a população I, embora as duas populações sejam constituídas pelos mesmos indivíduos. A heterogeneidade da população I é devida ao facto de ela ser espacialmente estratificada o que, do ponto de vista da amostragem, pode ser utilizado com vantagens.

Calculando novamente as médias por blocos (talhões) vê-se que, de facto, se reduziram grandemente as diferenças entre os blocos.

Coordenadas	**I**	**III**	**III**	**IV**
1-5	90	75	102	92
6-10	156	164	196	196
11-15	190	164	194	183
16-20	108	79	100	95

	I					II					III					IV					
20	130	24	153	--	153	--	112	18	200	--	106	35	100	--	147	12	118	--	165	--	
19	124	47	106	12	136	30	130	100	165	--	141	88	194	65	212	12	136	--	88	100	
18	177	24	165	--	136	--	124	30	171	12	106	88	82	6	177	47	147	82	165	118	**A**
17	165	30	112	53	124	53	118	30	153	30	118	--	224	24	136	35	118	65	159	141	
16	100	30	82	47	118	94	153	--	147	--	130	71	130	53	112	88	88	153	118	147	
15	224	--	247	--	217	6	230	71	130	59	259	82	277	141	100	118	147	171	171	200	
14	253	6	200	30	135	71	271	124	277	153	271	106	230	153	206	135	242	200	177	141	
13	212	71	277	53	265	136	212	118	206	106	171	88	289	165	259	183	183	277	247	194	**B**
12	224	94	283	159	247	171	300	188	100	71	318	124	277	71	306	141	177	271	200	177	
11	100	124	141	82	265	159	277	118	306	241	165	147	253	165	265	300	271	188	159	236	
10	277	165	330	77	253	188	218	118	177	106	353	88	330	118	253	177	171	177	194	241	
9	224	94	212	106	159	112	224	141	141	130	183	59	283	88	188	130	147	183	183	206	
8	271	141	318	59	200	224	271	106	218	194	253	153	260	242	200	165	147	77	259	253	**C**
7	277	200	277	218	206	124	236	159	230	100	230	212	294	124	165	94	294	159	212	259	
6	130	124	218	153	65	100	171	88	165	147	194	136	171	71	206	171	312	118	94	153	
5	218	71	130	82	118	100	130	124	82	159	171	100	147	212	124	124	177	94	183	159	
4	106	18	147	--	153	41	118	53	159	88	153	24	153	18	130	41	112	12	177	88	
3	130	35	200	41	194	30	100	35	141	--	165	--	153	18	147	--	177	35	194	106	**D**
2	77	30	165	--	159	--	159	--	183	6	118	6	124	18	124	100	94	--	159	71	
1	188	12	183	59	177	30	130	--	94	12	153	--	47	--	188	--	112	18	118	18	
	a	t	b	s	c	r	d	q	e	p	f	o	g	n	h	m	l	l	j	k	

Figura 2. População II – talhão de uma Mata Nacional - constituída por 400 parcelas de 0.1 ha cada, indicando-se em cada parcela o volume em m^3 ha^{-1}. A população II é composta das mesmas parcelas que a população I, mas a distribuição espacial foi alterada de modo a obter uma estrutura espacialmente menos estratificada.

Exemplo 3 – Avaliação do volume total numa mata de contornos irregulares (população III)

(adaptado de Loetsch *et al.*, 1977)

A figura 3 representa uma mata de contornos irregulares com área igual a 40 ha (população III) , tal como a população da figura 1. Os 400 indivíduos que constituem a população são os mesmos que constituem a população da figura 1, tendo-se apenas deslocado alguns grupos de parcelas para obter uma figura de contornos irregulares.

Esta população permite-nos avaliar melhor os problemas que se nos põem, na prática, para a definição dos indivíduos que constituem uma população.

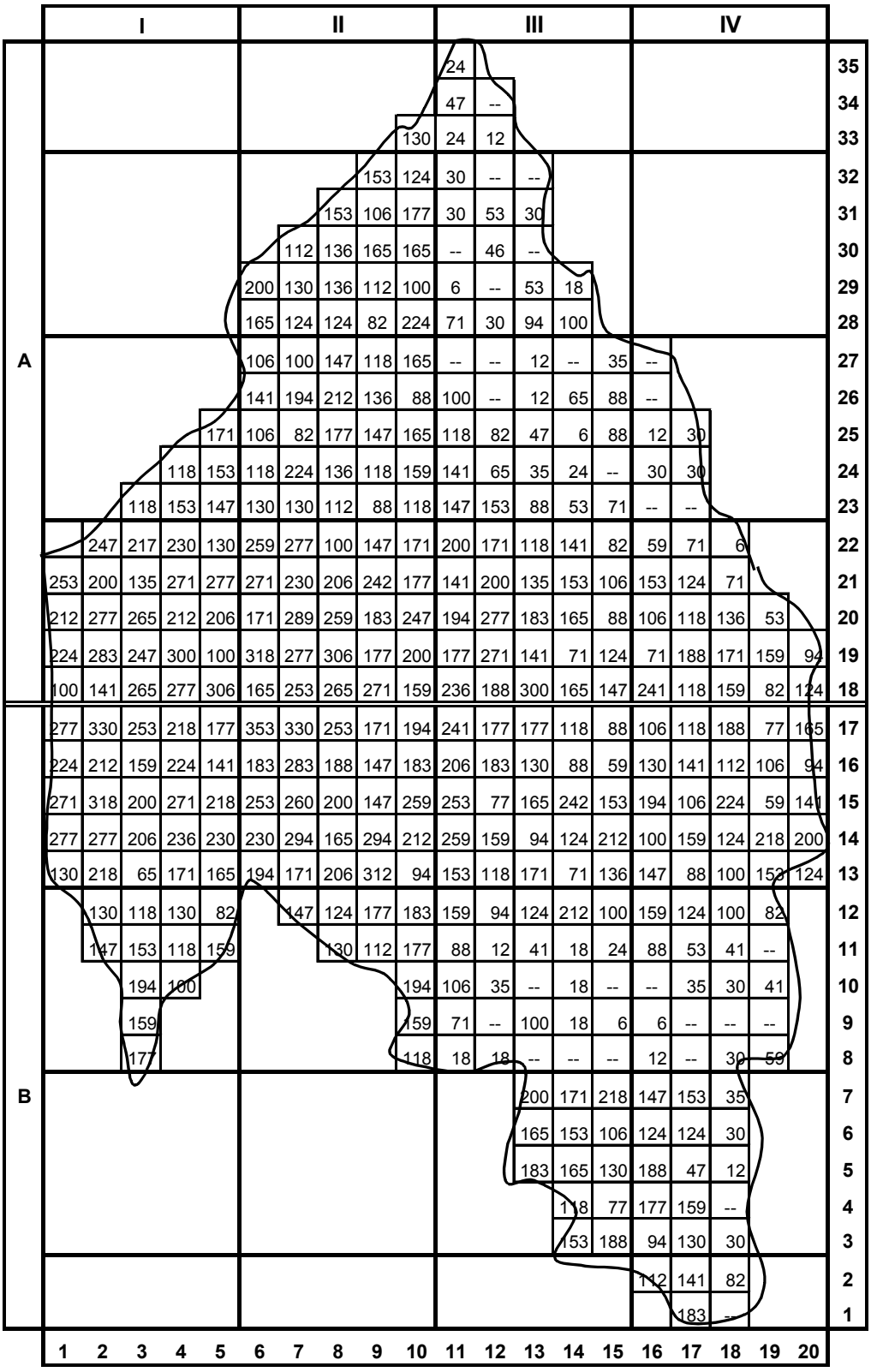


Figura 3. População III - mata de contornos irregulares - constituída por 400 parcelas de 0.1 ha cada, indicando-se em cada parcela o volume em $m^3 ha^{-1}$. A população III é constituída pelas mesmas parcelas que a população I.

Nesta população existe um grande número de parcelas que se encontram apenas parcialmente dentro dos limites da mata. Estas parcelas costumam designar-se por parcelas da bordadura do povoamento. À primeira vista, poder-se-ia pensar que estas parcelas poderiam ser excluídas da análise, uma vez que o volume total da mata se pode obter por multiplicação do volume médio por ha (avaliado com base em parcelas completas), multiplicado pela área total da mata.

Contudo esta solução pode originar estimativas enviesadas pelo facto de haver diferenças nítidas entre as zonas de bordadura dos povoamentos e o seu interior. O que fazer nestas circunstâncias? Uma das opções que vulgarmente se toma na prática é a de incluir dentro da população todas as parcelas que tenham mais de metade da sua área dentro dos limites da população e excluir aquelas em que esta condição não se verifique. Esta opção corresponde a admitir que a área que pertence à população e que é excluída é equivalente à área fora da população e que é contabilizada. Na altura da medição das parcelas de bordadura, o valor do volume (ou outra variável de interesse) encontrado na parte da parcela que pertence ao povoamento é depois extrapolado para a área da parcela completa, existindo diversos métodos para fazer esta extrapolação (vejam-se os apontamentos de medição de árvores e povoamentos).

Exemplo 4 – Volume total de uma espécie distribuída em pequenos povoamentos (população IV)

(adaptado de Loetsch *et al.*, 1977)

A população IV (figura 4) também foi obtida a partir da população I. Esta foi dividida em grupos de três a quinze malhas cada. Estes grupos foram então distribuidos numa matriz de 1600 malhas de tal maneira que o bloco AC contem as 125 malhas da coluna I e linhas 1-5 da coluna II da população I. O bloco AD contém as 75 malhas da linhas 6-20 da coluna II. O bloco BC contem as 150 malhas da coluna III e das linhas 1-10 da coluna IV e o bloco BD as 50 malhas das linhas 11-20 da coluna IV.

Deste modo, simulou-se uma distribuição espacial das parcelas que se encontra frequentemente nos inventários extensivos de grandes áreas (objectivo: avaliação de volumes para diversos estratos). A área desta população é de 160 ha.

Na figura 4, os indivíduos (malhas) que não estão preenchidos com qualquer valor representam áreas ocupadas por uma espécie diferente daquela sobre que incide o estudo ou áreas de clareira, enquanto que o valor inscrito nos indivíduos (malhas) que pertencem à espécie em questão representa o respectivo volume por ha.

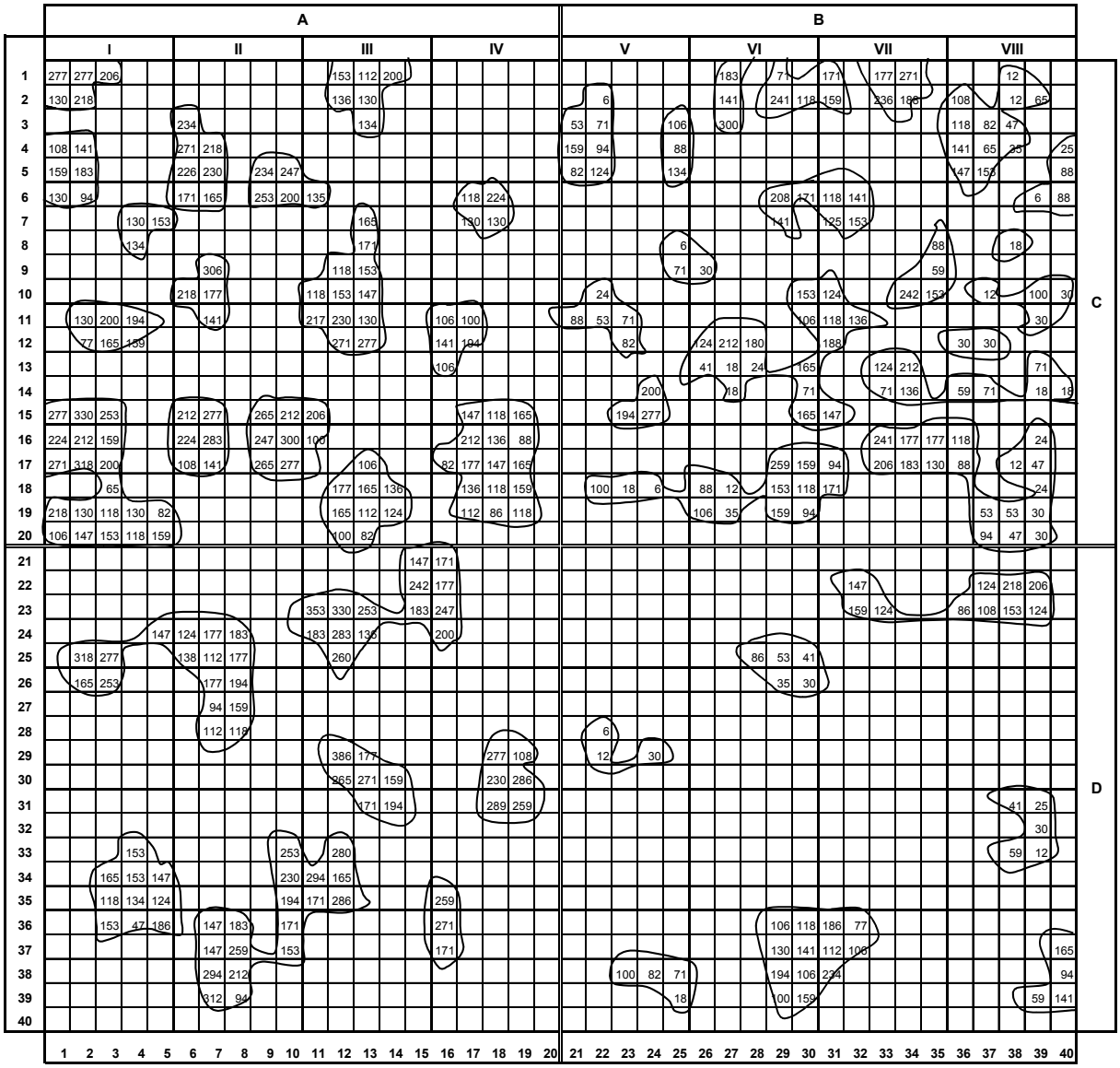


Figura 4. População IV, constituída por 1600 parcelas de 0.1 ha cada, indicando-se em cada parcela o volume em m^3 ha^{-1}. A população IV foi obtida a partir da população I distribuindo as 400 parcelas da população I numa matriz de 1600 parcelas, de acordo com a descrição do texto, pretendendo similar uma situação que se encontra frequentemente nos inventários das florestas tropicais.

Exemplo 5 – Avaliação do calibre médio e da porosidade média da cortiça numa pilha (população V)

(adaptado de Cumbre, 1999)

Suponha que pretende avaliar, por amostragem, o calibre e a porosidade médios da cortiça numa pilha com 100 x 8 x 2 m. Ambos os atributos são quantitativos. Podemos optar por definir cada uma das pranchas como indivíduo ou optar por considerar como indivíduos pequenos paralelepípedos com 1 metro de comprimento e largura e altura iguais às correspondentes medidas da pilha. No primeiro caso a população é de dimensão *N* desconhecida, no segundo a dimensão da população será igual ao comprimento da pilha de cortiça em metros. O segundo caso corresponde, como veremos, a uma amostragem por grupos. Uma outra alternativa, será a de, após seleccionar um paralelepípedo, não analisar todas as pranchas nele contidas, mas fazer uma segunda amostragem, ou seja, seleccionar um subconjunto das pranchas encontradas nesse paralelepípedo. Esta opção corresponde a uma amostragem por etapas, a qual não será objecto deste curso.

Exemplo 6 – Avaliação da distribuição da cortiça por classes de qualidade (1ª/3*, 4ª/5*, 6ª e refugo) numa parcela de 4 ha em montado de sobro (população VI)

O objectivo deste problema de amostragem é avaliar a distribuição da cortiça por classes de qualidade numa parcela de 4 ha. Trata-se neste caso da avaliação de um atributo qualitativo. Podemos definir como indivíduo um pequeno pedaço de cortiça de dimensão 20 x 20 cm. Podemos também utilizar as árvores como indivíduos ou, alternativamente, todas as árvores dentro de uma parcela de área fixa. Uma outra opção será ainda a de utilizar grupos com um número fixo de árvores. A parcela em questão foi exaustivamente medida (mais uma vez este problema, em termos de amostragem só tem interesse do ponto de vista académico) e encontra-se representada na figura 5. Os dados obtidos para cada árvore foram: perímetro à altura do peito, altura total, altura de descortiçamento, raios da copa, coordenadas das árvores, qualidade e calibre de um pedaço de cortiça ao nível do *d*, etc.

Exemplo 7 – Avaliação do preço médio da cortiça no campo (população VII)

Suponha que pretende avaliar o preço médio da cortiça de uma unidade de gestão, com uma área de 307.5 ha com base numa amostragem (figura 6). O preço da cortiça depende, quer do calibre, quer da qualidade. A tabela 1 dá indicação dos preços da cortiça para várias combinações de calibre e qualidade industrial, relativos ao preço da cortiça de maior valor (a qual apresenta, obviamente, um valor = 100). Trata-se neste caso novamente da avaliação de uma característica quantitativa. Os indivíduos podem definir-se de modo idêntico ao do exemplo anterior.

Tabela 1. Preços índices para cortiças de diferentes classes de qualidade (para indústria)

Calibre	Classe de qualidade			
	1ª/3ª	4ª/5ª	6ª	Refugo
14-18 mm	22	10	8	8
18-22 mm	31	13	8	8
22-27 mm	50	30	13	8
27-32 mm	100	60	28	8
32-40 mm	100	60	28	8
>40 mm	66	33	17	8

Exemplo 8 – Estudo da abundância de saca-rabos (*Herpestes ichneumon* L.) em Portugal

(adaptado de Borralho *et al.*, 1995)

Pretende-se avaliar a abundância de saca-rabos (*Herpestes ichneumon* L.) em Portugal. A abundância de uma espécie animal pode ser avaliada por diversos índices. Um deles é o número de indivíduos por unidade de área, outro a percentagem de ocorrência num determinado universo. No primeiro caso trata-se de um atributo quantitativo, no segundo um atributo qualitativo. As zonas de caça associativas (ou outra unidade territorial) são opções possíveis como indivíduos.

1.3 Caracterização de populações: parâmetros e distribuição de frequências

Geralmente as populações são constituídas por um grande número de indivíduos. Deste modo, o conhecimento exaustivo do atributo torna-se trabalhoso e dispendioso e, além disso, não permite uma "visão geral" da população. Os atributos exprimem-se então em termos de certos valores que se designam por parâmetros.

Os parâmetros facilitam a comparação entre diferentes populações e, em certos casos, sem eles a comparação seria impossível. Os parâmetros de localização permitem avaliar a ordem de grandeza do valor do atributo na população, os parâmetros de dispersão permitem avaliar a variabilidade da população.

Um outro elemento caracterizador das populações é a sua distribuição de frequências, ou distribuição dos indivíduos pelos diversos valores que o atributo pode tomar.

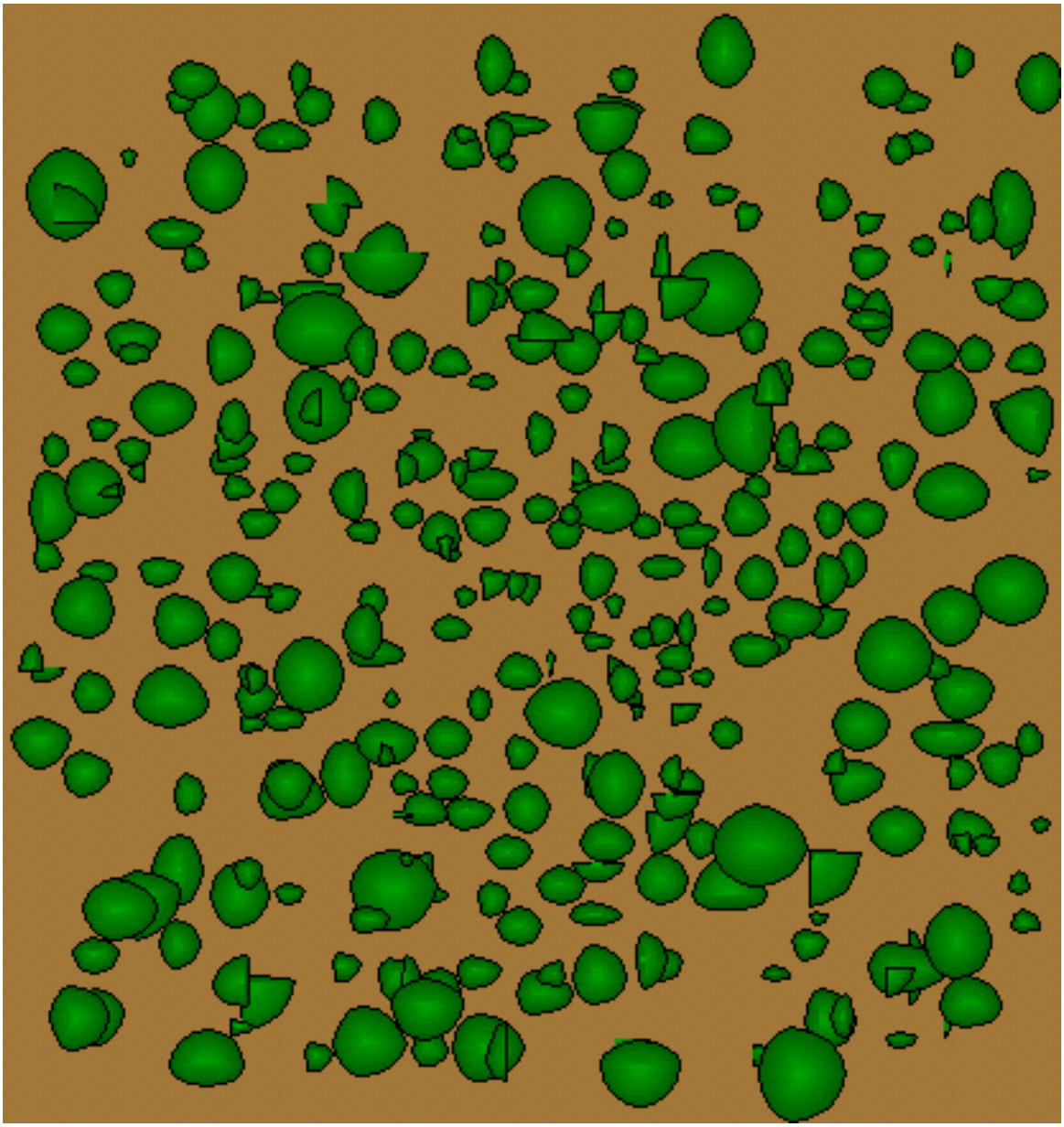


Figura 3. Mapa do povoamento de uma parcela de montado de sobro com 4 ha.

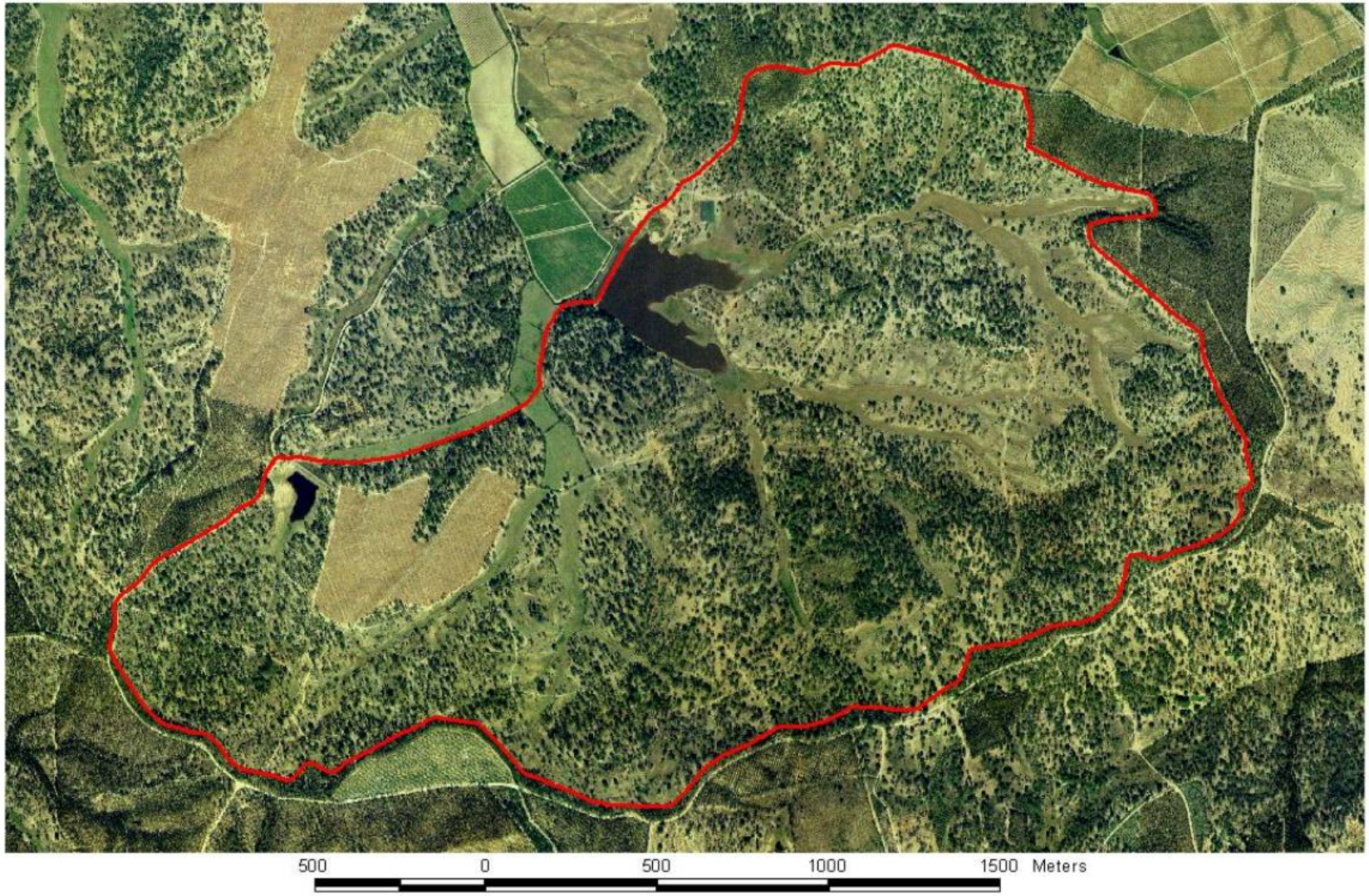


Figura 6. Ortofotomapa correspondente à população VII na qual se pretende avaliar o preço médio da cortiça.

1.3.1 Parâmetros de localização, dispersão e forma em populações de características quantitativas

1.3.1.1 A média e a variância

Sejam:

P – população de indivíduos u_i

N – dimensão da população

X – atributo em estudo

x_i – valor do atributo no indivíduo i

Os dois parâmetros mais importantes são o valor médio ou esperança matemática μ (parâmetro de localização) e a variância σ^2 (parâmetro de dispersão), calculados através das expressões:

Média: $$\mu = E(X) = \frac{\sum_{i=1}^{N} x_i}{N}$$

Variância: $$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Em alternativa à variância utiliza-se frequentemente como parâmetro de dispersão o quadrado médio S^2:

Quadrado médio: $$S^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N-1} = \frac{N-1}{N} \cdot \sigma^2$$

Pode utilizar-se a seguinte expressão para o cálculo do numerador da variância:

$$\sum_{i=1}^{N}(x_i-\mu)^2=\sum_{i=1}^{N}x_i^2-\frac{\left(\sum_{i=1}^{N}x_i\right)^2}{N}=\sum_{i=1}^{N}x_i^2-\bar{x}\cdot\sum_{i=1}^{N}x_i$$

Deste modo consegue-se, por um lado, uma maior facilidade no cálculo manual e, por outro lado, menores erros de cálculo ocasionados por um possível arredondamento da média.

Em vez de σ^2, utiliza-se frequentemente o desvio padrão $\sigma=\pm\sqrt{\sigma^2}$, para repor as unidades iniciais. Quando as populações diferem bastante no que respeita às médias, a comparação dos valores absolutos dos desvios padrões torna-se de pequena utilidade. Utiliza-se então o coeficiente de variação CV:

$$CV=100\cdot\frac{\sigma}{\mu}$$

Para além da média e da variância existem ainda outros parâmetros de localização e dispersão que, embora menos importantes, convém referir.

A título de exemplo, o valor destes parâmetros para os talhões de uma Mata Nacional dos exemplos 1 e 2 (populações I e II) encontra-se na tabela 2.

Tabela 2. Parâmetros do talhão de uma Mata Nacional (populações I e II)

Média	Variância	Quadrado médio	Desvio padrão	CV
136.44	6658.5408	6675.2288	81.60	59.81

1.3.1.2 Outros parâmetros de localização

Média geométrica: $N\cdot\sqrt{\prod_{i=1}^{N}x_i}$

Média harmónica: $1 \Big/ \left(\frac{1}{N} \sum_{i=1}^{N} \frac{1}{x_i} \right)$

Média quadrática: $\sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2}$

Mediana: Quando os indivíduos estão dispostos por ordem crescente, a mediana divide a população em duas partes iguais.

Quartis: Q_1, Q_2, Q_3 – idênticos à mediana, mas dividindo a população em quatro partes iguais.

Decis: $D_1, ..., D_9$ – idem, mas em 10 partes iguais.

Percentis: $P_1, ..., P_{99}$ – idem, mas em 100 partes iguais.

Nota: A mediana, os quartis, os decis e os percentis recebem o nome genérico de ***quantis***, sendo: Mediana = $Q_2 = D_5 = P_{50}$

Moda: Corresponde ao valor (valores) de maior frequência.

1.3.1.3 Outros parâmetros de dispersão

Amplitude do intervalo de variação: $Xmax - Xmin$

Amplitude interquartil: $Q_3 - Q_1$

Amplitude percentil: $P_{99} - P_1$

Desvio médio absoluto: $\frac{1}{N} \sum \left| x_i - \mu \right|$

1.3.1.4 Momentos

Momentos de ordem r: $$\mu_r' = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i^r$$

Momento central de ordem r, relativamente a μ: $$\mu_r = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \mu)^r$$

Momento central de ordem r, relativamente a α: $$\mu_r(a) = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \alpha)^r$$

Note-se que a média é o momento de ordem 1, enquanto que a variância é o momento central de ordem 2.

1.3.1.5 Parâmetros de forma

Com o auxílio dos momentos, definem-se ainda os parâmetros de forma:

Coeficiente de assimetria (skewness): $$c_1 = \frac{\mu_3}{\sigma^3} = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \mu)^3 \cdot \frac{1}{\sigma^3}$$

A figura 6 exemplifica as formas das funções densidade de probabilidade quando a distribuição é simétrica - coeficiente de assimetria nulo - e quando é assimétrica à direita ou negativa ou assimétrica à esquerda ou positiva – coeficientes de assimetria, respectivamente, menor e maior que zero.

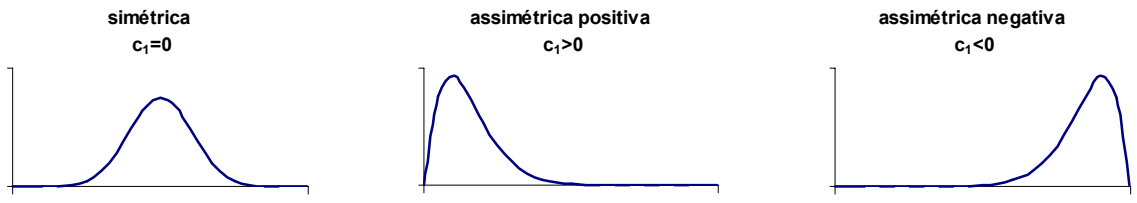


Figura 6. Representação gráfica de distribuições simétrica, assimétrica à esquerda e assimétrica à direita.

Coeficiente de achatamento: $$c_2 = \frac{\mu_4}{\sigma^4} = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \mu)^4 \cdot \frac{1}{\sigma^4}$$

A figura 7 ilustra as formas das funções densidade de probabilidade quando a distribuição é mesocúrtica – coeficiente de achatamento igual a 3 –, platicúrtica ou achatada – coeficiente de achatamento menor que 3 – ou leptocúrtica ou abicada – coeficiente de achatamento maior do que 3.

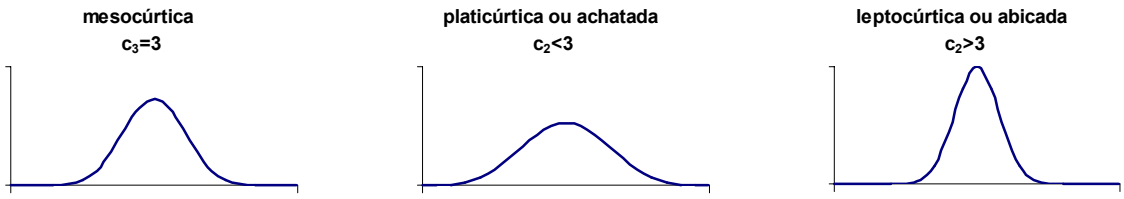


Figura 7. Representação gráfica de distribuições mesocúrtica, platicúrtica e leptocúrtica.

Por vezes utiliza-se um coeficiente alternativo a c_2, o qual é nulo para as distribuições mesocúrticas, portanto para a distribuição normal:

$c_1 = c_2 - 3$,

1.3.1.6 Cálculo dos parâmetros com os dados agrupados

Muitas vezes faz-se, para facilidade de cálculo (nomeadamente do cálculo manual), o agrupamento dos dados em classes e posterior cálculo dos parâmetros com base nos valores das classes e frequências respectivas. Apresentam-se em seguida, a título de exemplo, as expressões para o cálculo da média e da variância com os dados agrupados.

Sejam:

M – número de classes

X_j – valor central da classe (j=1, ..., M)

N_j – frequência da classe j

Temos então:

Média: $$\mu = \frac{\sum_{j=1}^{M} N_j x_j}{\sum_{j=1}^{M} N_j}$$

Variância: $$\sigma^2 = \frac{\sum_{j=1}^{M} N_j (x_j - \mu)^2}{\sum_{j=1}^{M} N_j}$$

$$\sum_{j=1}^{M} N_j (x_j - \mu)^2 = \sum N_j x_j^2 - \frac{\left(\sum_{j=1}^{M} N_j x_j \right)^2}{\sum_{j=1}^{M} N_j}$$

1.3.2 Parâmetros de localização e dispersão em populações com atributos qualitativos

Se o atributo em estudo é qualitativo, os parâmetros calculam-se do mesmo modo:

Sejam:

N – número total de indivíduos

A – número de indivíduos que possuem a característica C

$p = \frac{A}{N}$ - proporção de C na população

Seja *X* a variável indicatriz da característica *C*, isto é, *X* toma o valor 1 ou 0 conforme o indivíduo possui ou não a característica *C*. Então:

Média: $$\mu = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i = \frac{A}{N} = p$$

Variância: $$\sigma^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - p)^2 \equiv \frac{1}{N} \cdot \sum_{i=1}^{N} \left(x_i^2 - p \cdot \sum_{i=1}^{N} x_i \right) =$$

$$= \frac{1}{N} \cdot \sum_{i=1}^{N} x_i^2 - p^2 =$$

$$= \frac{1}{N} \cdot N \cdot p - p^2 = p - p^2 = p \cdot (1 - p) = p \cdot q$$

1.3.3 Distribuições de frequências

A distribuição de frequências de uma população é uma representação "esquemática", por vezes gráfica, da função de distribuição da variável ou atributo em questão. Para além de poder ser representada com base nas frequências acumuladas, também o pode ser com recurso a frequências absolutas ou relativas.

Para o estudo das distribuições de frequências convém distinguir entre atributos discretos e contínuos. Os primeiros tomam valores num conjunto cujos elementos se podem pôr em correspondência com o conjunto dos naturais, os segundos correspondem a um intervalo da recta real. Os atributos discretos correspondem, geralmente, a atributos qualitativos e os atributos contínuos a atributos quantitativos.

No caso de um atributo ou variável discreta a distribuição de frequências fica definida pelo conjunto de valores que a variável pode tomar e pelas frequências respectivas, geralmente organizadas numa tabela de frequências.

Sejam:

f – frequência absoluta

f_r – frequência relativa

f_a - frequência acumulada

A tabela 3 representa, esquemáticamente, uma distribuição de frequências de uma variável discreta, indicando as frequências absolutas, relativas e acumuladas

Tabela 3. Distribuição de frequências de uma variável discreta

X	F	f_r	f_a
X_1	f_1	f_1/N	f_1
x_2	f_2	f_2/N	$f_1 + f_2$
x_3	f_3	f_3/N	$f_1 + f_2 + f_3$
.	.	.	.
.	.	.	.
.	.	.	.
x_k	f_k	f_k/N	$f_1 + ... + f_k$

Uma tabela de frequências pode representar-se graficamente em polígonos de frequência ou em histogramas. A figura 8 mostra a distribuição de frequências da população VI (qualidade da cortiça avaliada, em cada árvore, num pedaço de cortiça retirado a 1.30 m) representada em histograma e em polígono de frequências

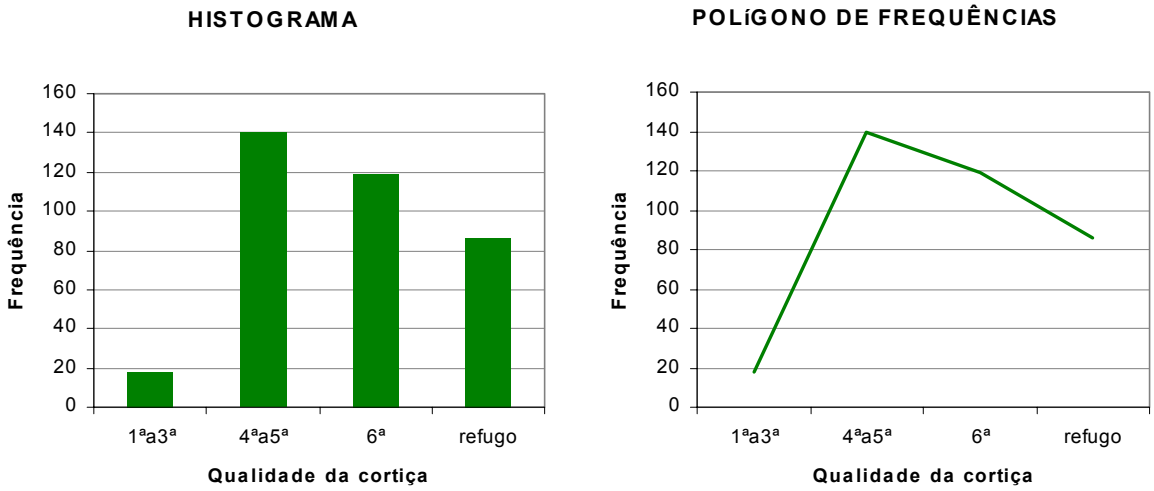


Figura 8. Distribuição de frequências da população VI

No caso de uma variável contínua há que construir uma tabela de dados agrupados em classes. Para construir estas tabelas, a maior dificuldade consiste em encontrar qual a amplitude das classes a utilizar, pois esta influencia grandemente a forma do polígono: é frequente um polígono mostrar mais do que um pico como consequência da pequena amplitude das classes; por outro lado, se a dimensão das classes é demasiado grande as características da distribuição de frequências diluem-se dentro das classes. A figura 9 representa a distribuição de frequências (histograma e polígono de frequências) das populações I e II com base em classes de volume com amplitude de 30 m^3.

Algumas regras práticas:

- número de classes deve andar entre 8 e 20
- número de classes não deve ser menor que $\sqrt[3]{N}$ e, de preferência, deve ser maior
- sendo α a amplitude das classes e *n o seu número, deve verificar-se:*

$$\alpha = \log_2(n+1) = 3.322\log_{10}(n+1)$$

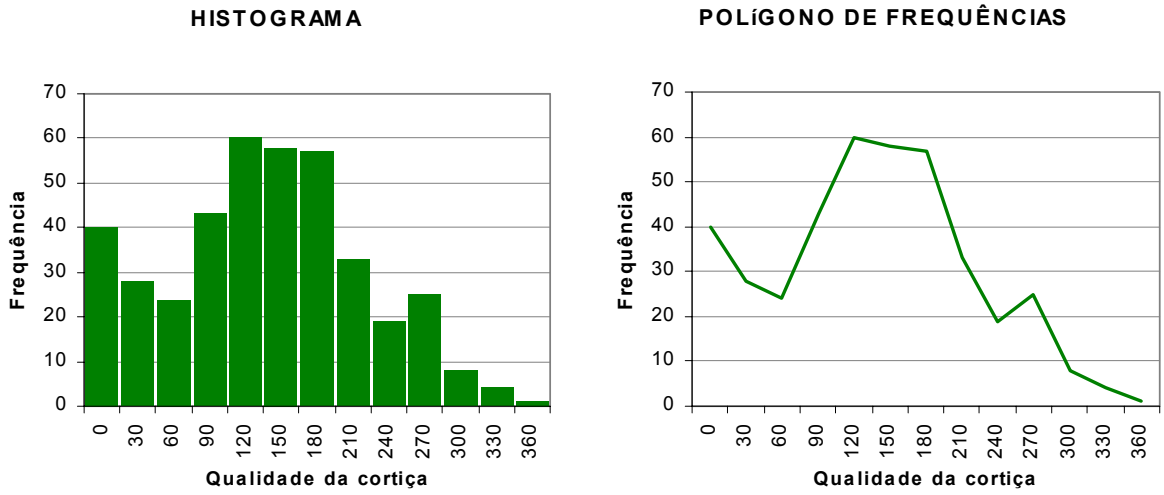


Figura 9. Distribuição de frequências das populações I e II, classes de 30 m^3

1.4 Atributos e variáveis aleatórias

Seja $U = \{u_1, u_2, ..., u_N\}$ uma população ou um universo com *N* indivíduos. Vamos ver que qualquer atributo ou característica mensurável sobre esta população se pode interpretar como uma variável aleatória. Podemos então falar indiscriminadamente em população *X*, característica mensurável *X*, atributo *X* ou variável aleatória *X*.

Pensemos característica mensurável *X* e designemos $X(u_i)$, o valor do atributo *X* no indivíduo u_i, por y_i. *X* pode então tomar *N* valores não necessariamente distintos:

$$y_1, y_2, y_3, ..., y_N$$

Vamos admitir que há apenas *k* y_i distintos e representemo-los ordenados por ordem crescente:

$$x_1, x_2, x_3, ..., x_k$$

Em *U* podemos introduzir uma relação de equivalência ρ:

$$u_i \ \rho \ u_j \ \Leftrightarrow \ X(u_i) = X(u_j)$$

Obtemos então o seguinte conjunto quociente de *k* elementos:

$$U/\rho = \{C_1, C_2, ..., C_k\}$$

com $C_i = \{u_i \mid X(u_i) = x_i\}$

Seja $f_i = \#\ (C_i)$ a dimensão da classe C_i, ou seja, a frequência absoluta dos indivíduos que tomam o valor x_i.

Assim:

$$\sum_{i=1}^{k} f_1 = N$$

onde $\frac{f_1}{N}$ representa a frequência relativa ou a proporção de indivíduos na classe i

e, portanto, $\sum_{i=1}^{k} \frac{f_i}{N} = 1$

Consideremos agora a experiência aleatória **E** "retirar ao acaso um indivíduo da população e repô-lo" (a reposição é indispensável para poder repetir indefinidamente a experiência sob as mesmas condições, característica das experiências aleatórias); posso associar-lhe uma variável aleatória *X* com valores $x_1, x_2, x_3, ..., x_k$ e probabilidades $p_1 = f_1/N_1, ..., p_k = f_k/N_k$:

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_k \\ p_1 = \frac{f_1}{N_1} & p_2 = \frac{f_2}{N_2} & & p_k = \frac{f_k}{N_k} \end{pmatrix}$$

Podemos agora calcular *E*(*X*) e *v*(*X*) e facilmente chegamos às seguintes conclusões:

$$E(x) = \frac{1}{N} \cdot \sum_{i=1}^{N} y_i = \bar{y}$$

$$v(x) = \frac{1}{N} \cdot \sum (y_i - \bar{y})^2 = \sigma_y^2$$

Fica assim estabelecida a ligação entre a teoria (modelos teóricos das variáveis aleatórias) e os problemas práticos com que temos que lidar.

1.5 Distribuições teóricas de variáveis aleatórias

1.5.1 Modelos das variáveis aleatórias mais importantes para a amostragem

Tal como ficou dito, pode sempre estabelecer-se a correspondência entre uma característica mensurável *X* e uma variável aleatória, usualmente também designada por *X*. Compreende-se assim facilmente a importância que têm para a amostragem os modelos teóricos das variáveis aleatórias, salientando-se a importância das seguintes distribuições: binomial, normal, *t*-student, χ^2 e *F*-Snedecor.

Aconselham-se portanto os alunos a reverem a matéria leccionada na Estatística sobre variáveis aleatórias e respectivos modelos!!...

1.5.2 Teoremas sobre distribuições

Convém rever alguns teoremas sobre variáveis aleatórias, bastante importantes para a compreensão da matéria da amostragem, de que tratam os capítulos posteriores.

1 – Se X_1 e X_2 são variáveis aleatórias independentes e normais com valores médios μ_1 e μ_2 e variâncias σ_1^2 e σ_2^2, então a variável aleatória $X=X_1+X_2$ é ainda normal com valor médio $\mu_1+\mu_2$ e variância $\sigma_1^2+\sigma_2^2$.

$$\left.\begin{array}{l} X_1 \cap \mathrm{N}\ (\mu_1,\sigma_1^2) \\ \text{Independentes} \\ X_2 \cap \mathrm{N}\ (\mu_2,\sigma_2^2) \end{array}\right| \Rightarrow X = X_1 + X_2 \cap \mathrm{N}\ (\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$$

2 - Para $n \geq 30$, Fisher demonstrou que, sendo χ_n^2 um qui-quadrado com *n g.l.*, a v.a.

$$z = \sqrt{2\chi_n^2} - \sqrt{2n-1} \cap N(0,1)$$

é aproximadamente normal reduzida, sendo a aproximação tanto melhor quanto maior for *n*.

3 - A distribuição χ^2 com *n g.l.* tende para uma distribuição normal (*n, 2n*) quando $n \to \infty$

$$\chi_n^2 \xrightarrow[n\to\infty]{} N(n,2n)$$

4 - Se X_1 e X_2 são v.a. independentes com distribuição χ^2, respectivamente com n_1 e n_2 *g.l.*, então a v.a. $X=X_1+X_2$ tem uma distribuição χ^2 com $n = n_1+n_2$ *g.l.*:

$$\left.\begin{array}{l} X_1 \cap \chi_{n_1}^2 \\ \text{independentes} \\ X_2 \cap \chi_{n_2}^2 \end{array}\right| \Rightarrow X = X_1 + X_2 \cap \chi_{n_1+n_2}^2$$

5 - Teorema limite central

Seja $X_1, X_2, ..., X_n$ uma sequência de v.a. independentes, com $E(X_i)=\mu_i$ e $v(x_i) = \sigma_i^2$ *(i=1, 2, ...,n)*.

Sob determinadas condições gerais (quase sempre satisfeitas na prática) a variável

$$X=X_1+X_2+...+X_n$$

tem uma distribuição assimptoticamente normal de parâmetros.

$$\mu = \sum_{i=1}^{n} \mu_i$$

e

$$\sigma = \sum_{i=1}^{n} \sigma_i^2$$

Consequentemente, a variável

$$z_n = \frac{X - \mu}{\sigma}$$

tem uma distribuição assimptoticamente normal reduzida $N(0,1)$.

6 - Se $X \cap N(\mu,\sigma^2)$ e *(x_1, x_2, ...,x_n)* é uma amostra de dimensão *n* de *X*, então

$$t = \sqrt{n-1}\,\frac{\overline{X_n} - \mu}{s}, \qquad \left(\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} x_i\right), \left(s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\right)$$

tem uma distribuição de *Student* com *(n-1)* graus de liberdade.

Se $X \cap N(\mu,\sigma^2)$, então $\overline{X_n} \cap N\left(\mu,\sigma^2/n\right)$ e $\sqrt{n}\,\frac{\overline{X_n} - \mu}{\sigma} \cap N(0,1)$

Ora $\frac{ns^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2$,

pelo que

$\frac{X - \mu}{\sigma} \cap N(0,1)$, pelo que $\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$ tem uma distribuição χ^2 com *n g.l.* (recorde-se que a amostra é independente)

$$\sum (x_i - \mu)^2 = \sum (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 + 2(\bar{x} - \mu)\underbrace{\sum (x_i - \bar{x})}_{0}$$

$$= \sum (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

Portanto

$$\sum \left(\frac{x_i - \mu}{\sigma} \right)^2 = \frac{ns^2}{\sigma^2} + n\left(\frac{\bar{x} - \mu}{\sigma} \right)^2$$

$$\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right)^2 \cap \chi_1^2$$

Temos assim:

$$\chi_n^2 = \frac{ns^2}{\sigma^2} + \chi_1^2 \Rightarrow \frac{ns^2}{\sigma^2} \cap \chi_{n-1}^2$$

Para obter a variável *t* com *(n-1) g.l.* basta dividir uma v.a. $N(0,1)$ pela $\sqrt{}$ de um $\frac{\chi_{n-1}^2}{n-1}$, ou seja

$$\frac{\frac{\overline{X} - \mu}{\sigma}\sqrt{n}}{\frac{s}{\sigma}\frac{\sqrt{n}}{\sqrt{n-1}}} = \sqrt{n-1}\frac{\overline{X} - \mu}{s} \cap t_{n-1}$$

Uma consequência deste teorema é que $\sqrt{n}\frac{\overline{X} - \mu}{s_c} \cap t_{n-1}$

7 - Sejam *X* e *Y* variáveis aleatórias

$$X \cap N(\mu_1, \sigma_1^2)$$

$$Y \cap N(\mu_2, \sigma_2^2)$$

sejam $(x_1, x_2, \ldots, x_n)$ e $(y_1, y_2, \ldots, y_n)$ amostras independentes de *X* e de *Y*, respectivamente.

Então

$$\frac{n_1 s_1^2}{\sigma_1^2} \cap \chi^2_{n_1-1}$$

$$\frac{n_2 s_2^2}{\sigma_2^2} \cap \chi^2_{n_2-1}$$

O quociente $\frac{\chi^2_{n_1-1}/(n_1-1)}{\chi^2_{n_2-1}/(n_2-1)}$ tem uma distribuição F com *(n_1-1)* e *(n_2-1) g.l.*

Assim

$$\frac{n_1 s_1^2/\sigma_1^2}{n_2 s_2^2/\sigma_2^2} \cdot \frac{n_2-1}{n_1-1} = \frac{n_1(n_2-1)}{n_2(n_1-1)} \cdot \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{s_1^2}{s_2^2} \cap F_{n_1-1;n_2-2}$$

1.6 Exercícios

Os dados relativos às populações I a IV e VI encontram-se no directório D:\partilha\Inventário Florestal\textos\amostragem dos computadores do Laboratório de Informática do DEF nos seguintes ficheiros:

Populações.xls figuras 1 a 4 em EXCEL (o ficheiro contém ainda a população I.A)

POP-I.xls população I sob forma tabular, com os seguintes dados nas colunas A a E: linha-talhão, coluna-talhão, linha, coluna, volume

POP-I.b.xls população I.b sob forma tabular, com os seguintes dados nas colunas A a F: linha-talhão, coluna-talhão, linha, coluna, estrato, volume

POP-II.xls população II sob forma tabular, com os seguintes dados nas colunas A a E: linha-talhão, coluna-talhão, linha, coluna, volume

POP-III.xls população III sob forma tabular, com os seguintes dados nas colunas A a E: linha-talhão, coluna-talhão, linha, coluna, volume

POP-IV.xls população IV sob forma tabular, com os seguintes dados nas colunas A a F: linha-talhão, coluna-talhão, linha, coluna, volume, espécie (variável qualitativa: presença=1, ausência=0).

POP-VI.xls população VI sob forma tabular, com os seguintes dados nas colunas A a O: arvore, dapsc, coordenada x, coordenada y, altura de bifurcação, altura de descortiçamento, altura da base da copa, altura total, raios da copa segundo os azimutes 300, 30 120 e 210, saúde, calibre da cortiça, qualidade da cortiça.

Em relação às populações V e VII dispomos apenas dos dados obtidos em amostragens realizadas sobre essas populações, os quais se encontram nos ficheiros:

POP-V.xls população V sob forma tabular, com os seguintes dados nas colunas A a E: comprimento da pilha, situação na largura, situação na altura, calibre, qualidade

POP-VII.xls população VII sob forma tabular, com os seguintes dados nas colunas A a H: parcela, estrato, cala nº, qualidade, calibre, direcção, distância ao centro da parcela, preço/@

Estes ficheiros devem ser utilizados nos exercícios que se seguem.

1.6.1 Noção de população e sua caracterização

Parâmetros

Ex1.1 Demonstre as seguintes propriedades da média aritmética.

a) A soma dos desvios para a média é zero

$$\sum_{i=1}^{N}(x_i - \mu) = 0$$

b) A soma dos quadrados dos desvios para a média é mínima

$$\sum_{i=1}^{N}(x_i - \mu)^2 = min$$

Ex1.2 Verifique a expressão

$$\sum_{i=1}^{N}(x_i - \mu)^2 = \sum x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2$$

Ex1.3 Calcule alguns parâmetros da população I: média, variância, desvio padrão e coeficiente de variação. Não se esqueça que os dados correspondentes a esta população se encontram disponíveis nos computadores do Laboratório de Informática do DEF no ficheiro POP-I.xls.

Ex1.4 O mesmo que **Ex1.3**, mas em relação à população II (ficheiro POP-II.xls).

Ex1.5 O mesmo que **Ex1.3**, mas trabalhando com os dados agrupados em classes de 30 m^3 (classes: 0; 1-30; 31-60; ...)

Ex1.6 O mesmo que **Ex1.3**, mas calculando o valor dos parâmetros em cada talhão. Calcule também estes parâmetros para a população II.

1.6.2 Distribuição de frequências

Ex1.7 Faça os polígonos de frequência para a população Ib e para os seus três estratos. Utilize as frequências relativas, absolutas e acumuladas (relativas). Procure interpretar os resultados.

Sugestão: Utilize classes de amplitude = 30 m^3ha^{-1}, mas procure justificar este número.

Ex1.8 Para a mesma população, calcule para cada estrato qual a percentagem de indivíduos que estão no intervalo $\mu_j \pm \sigma_j$ (esta percentagem depende do tipo de distribuição de frequência).

Sugestão: Utilize a curva das frequências acumuladas.

Ex1.9 Faça a distribuição de frequências dos diâmetros (classes de 5 cm de amplitude, sendo a 1ª classe com o limite inferior igual a 2.5 cm) da população VI.

Calcule ainda os seguintes parâmetros de localização, dispersão e forma:

- Média
- Variância
- Coeficiente de assimetria
- Coeficiente de achatamento

1.6.3 Modelos de variavéis aleatórias mais importantes (revisão)

Distribuição normal

Ex1.10 Mostre que os pontos de abcissas $X = \mu \pm \sigma$ são pontos de inflexão da curva normal $\mathcal{N}(\mu, \sigma^2)$

Ex1.11 Suponha a variável aleatória $X \sim \mathcal{N}(2, 0.16)$. Utilizando uma tabela da distribuição normal, assim como a função NORMSINV do EXCEL, calcule as seguintes probabilidades.

a) $P(X \geq 2.3)$

b) $P(1.8 \leq X \leq 2.1)$

Ex1.12 Admita que o diâmetro das árvores de um povoamento irregular de pinheiro bravo está distribuido normalmente com média 20 cm e variância 105.0625. Qual é a probabilidade de que o diâmetro exceda os 40 cm?

Suponha que uma árvore deste povoamento é seleccionada para corte final se o seu diâmetro diferir da média mais do que 10 cm. Qual é a probabilidade de que uma árvore seja abatida para corte final?

Ex1.13 Os erros de um aparelho de medição de alturas de árvores têm distribuição normal com valor médio nulo e desvio padrão igual a 1 m. Qual é a probabilidade de obter um erro de medição superior a 1 m? e a 2 m? e a 3 m?

Ex1.14 Suponha que tem dois povoamentos florestais, P1 e P2, com distribuições de diâmetros $\mathcal{N}(40, 36)$ e $\mathcal{N}(45, 9)$, respectivamente. Qual o povoamento que deve ser preferido para exploração, se o objectivo for abater árvores com um diâmetro superior a 45 cm? E se este diâmetro for de 48 horas?

Ex1.15 Suponha que X é $\mathcal{N}(\mu,\sigma^2)$. Determine c (função de μ e σ) tal que $P(X \leq c)=2\ P(X>c)$. Recorra a uma tabela da distribuição normal ou à função NORMSINV do EXCEL.

Ex1.16 Recorrendo à função NORSMINV do EXCEL, elabore uma tabela da distribuição normal, ou seja, calcule os valores críticos da distribuição normal correspondentes a diferentes valores de $P(X<z_{alfa/2})$.

Distribuição de χ^2

Ex1.17 Considere uma variável aleatória $X \sim \chi^2_{10}$ (qui-quadrado com 10 graus de liberdade). Se quizermos encontrar dois números a e b tais que $P(a < X < b)=0.90$, verifica-se que existem diversos pares (a , b) nestas condições.

a) Encontre dois pares de valores (a, b) que satisfaçam a condição acima expressa. Utilize uma tabela de distribuição de χ^2 ou a função CHIINV do EXCEL.

b) Suponha que impomos a restrição adicional $P(X < a)=P(X > b)$. Quantos pares (a , b) existem nestas condições?

Ex1.18 Suponha a variável aleatória $X \sim \mathcal{N}(0, 25)$. Calcule $P(1 < X^2 <4)$

Distribuição t-Student

Ex1.19 Recorrendo à função TINV do EXCEL, elabore uma tabela t-Student, ou seja, calcule os valores críticos da distribuição t-Student correspondentes a diferentes valores de $P(X<t_{alfa/2})$.

2 Amostragem e amostra

2.1 Definição

A amostragem pode definir-se como a análise de uma parte da população com o objectivo de tirar conclusões (inferir) para todo o conjunto. Ao subconjunto observado dá-se o nome de amostra.

A utilização de amostragem no estudo de uma população justifica-se sempre que a dimensão da população seja excessivamente grande, tornando-se impossível observar (medir) todos os indivíduos. Há ainda casos em que a medição da variável é destrutiva (p.e. determinação de biomassas); neste caso o estudo da população só pode ser feito por amostragem. Mesmo quando é possível observar todos os indivíduos, a rapidez e redução de custos jogam a favor da utilização de amostragem. Note-se ainda que em muitos casos ao estudo de uma população por amostragem estão associadas maior profundidade e precisão na avaliação da variável de interesse do que ao correspondente estudo por enumeração total.

2.2 Representatividade da amostra

Um problema importante que se põe é o da representatividade da amostra, isto é, os indivíduos que constituem a amostra devem dar uma imagem tão correcta quanto possível da população. Uma amostra será representativa se a sua estrutura for semelhante à da população que lhe deu origem, ou seja, se os diversos tipos de indivíduos (definidos de acordo com os valores do atributo em questão) estiverem representados em proporções semelhantes aquelas com que aparecem na população.

2.3 Métodos para a selecção dos indivíduos que fazem parte da amostra

A selecção dos indivíduos que vão ser objecto de análise (amostra) pode ser realizada por três processos:

- Amostragem aleatória
- Amostragem selectiva
- Amostragem sistemática

2.3.1 Amostragem aleatória

Numa amostragem aleatória a tiragem dos indivíduos é feita com probabilidades conhecidas (não necessariamente iguais) e a amostra é obtida através de um mecanismo aleatório, geralmente uma tabela de números aleatórios ou algoritmos geradores de números pseudo-aleatórios (ver 2.3). Como consequência, cada amostra possível tem uma probabilidade de selecção conhecida pelo que, para qualquer processo de amostragem, nestas condições, é possível calcular a distribuição de frequências para os estimadores que este processo origina. Toda a teoria da amostragem se baseia no pressuposto de que a amostra foi obtida por um mecanismo aleatório.

2.3.1.1 Sorteio com e sem reposição.

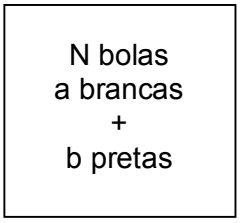
Para sortear (obter casualmente) os indivíduos que fazem parte da amostra utiliza-se um modelo de urna de Bernoulli ou qualquer outro processo equivalente, como por exemplo, uma tabela de números casuais ou aleatórios. Uma tabela de números aleatórios é um conjunto grande de números (várias páginas) nos quais há a garantia de que, se n é grande, então todos os números têm a mesma probabilidade de ser seleccionados. Geralmente utiliza-se este último processo por razões de ordem prática, mas os dois métodos são equivalentes. Hoje em dia é frequente recorrer-se a números pseudo-aleatórios gerados em máquinas de calcular ou programas de computador com base em algoritmos que tentam reproduzir a característica das tabelas de números aleatórios, ou seja, que originam séries de números em que todos têm a mesma probabilidade de ser sorteados. Os números pseudo-aleatórios gerados em máquina de calcular ou programa de computador variam geralmente entre 0 e 1, sendo posteriormente transformados pelo utilizador em números entre 1 e n do seguinte modo: multiplica-se o número aleatório ente 0 e 1 por n, toma-se a parte inteira e adiciona-se 1. Garante-se deste modo que todos os números entre 1 e n têm a mesma probabilidade de ser sorteados.

Já vimos (1.2) que as tiragens se devem fazer com reposição para garantir que todos os indivíduos têm a mesma probabilidade de serem sorteados e, assim, que a experiência aleatória “retirar ao acaso um indivíduo e repô-lo” se repita indefinidamente sob as mesmas condições. Nestas condições a amostra é independente, ou seja, pode ser interpretada como um valor do vector aleatório $(X_1, X_2, ..., X_n)$ de componentes independentes.

Na prática, contudo, é preferível fazer tiragens sem reposição para evitar trabalhar realmente com uma amostra de grandeza (n-1) quando pretendíamos uma amostra de grandeza n. Se a população for muito grande ($N \to \infty$), a probabilidade de um indivíduo sair duas vezes é muito pequena (no caso contínuo é mesmo nula), pelo que é praticamente equivalente utilizar os dois tipos de tiragens. No entanto se a população for pequena há que ter em conta o carácter finito da população no caso das tiragens serem feitas sem reposição.

Analisemos melhor o problema das tiragens com e sem reposição recorrendo a alguns exemplos.

Exemplo1 - Consideremos uma urna com N bolas, sendo a bolas brancas e b bolas pretas. Pretendemos calcular as distribuições conjuntas correspondentes a uma amostra de duas bolas, com e sem reposição, assim como as correspondentes distribuições marginais.



bola branca $\Leftrightarrow$ **acontecimento** A

bola preta $\Leftrightarrow$ **acontecimento** $\overline{A}$

Com reposição

$\frac{a}{N} \to$ probabilidade de sair bola branca em qualquer tiragem

$\frac{a}{N}\ \frac{a}{N} \to$ probabilidade de saírem duas bolas brancas em duas tiragens

$\left(\frac{a}{N}\right)^k \to$ probabilidade de saírem k bolas brancas em k tiragens

Para simplificar, consideremos apenas duas tiragens (x_1, x_2), valor observado do vector aleatório(X_1, X_2), com a distribuição conjunta e as correspondentes distribuições marginais.

$$(X_1, X_2)\begin{vmatrix}(1,1) & (1,0) & (0,1) & (0,0)\\ \frac{a^2}{N^2} & \frac{ab}{N^2} & \frac{ab}{N^2} & \frac{b^2}{N^2}\end{vmatrix}$$

$$X_1\begin{vmatrix}0 & 1\\ \frac{b}{N} & \frac{a}{N}\end{vmatrix} \quad \text{e} \quad X_2\begin{vmatrix}0 & 1\\ \frac{b}{N} & \frac{a}{N}\end{vmatrix}$$

Como as distribuições marginais de X_1 e X_2 são iguais, as v.a. são semelhantes (identicamente distribuídas), além de independentes (a independência é consequência da reposição).

Resumindo, se a amostragem é feita com reposição (X_1, X_2) é um par aleatório de componentes independentes e semelhantes.

Sem reposição

Vamos ver que, embora se perca a independência, as v.a. continuam a ser semelhantes, ou seja, identicamente distribuídas.

Vejamos qual é, neste caso, a distribuição conjunta do par aleatório (X_1, X_2):

$$(X_1, X_2)\begin{vmatrix}(1,1) & (1,0) & (0,1) & (0,0)\\ \frac{a}{N}\cdot\frac{a-1}{N-1} & \frac{a}{N}\cdot\frac{b}{N-1} & \frac{b}{N}\cdot\frac{a}{N-1} & \frac{b}{N}\cdot\frac{b-1}{N-1}\end{vmatrix}$$

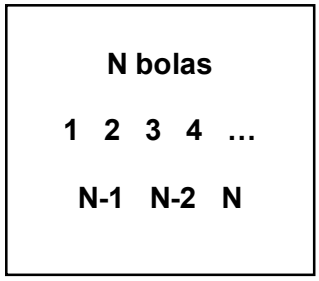
As distribuições marginais respectivas são

$$X_1\begin{vmatrix}1 & 0\\ \frac{a}{N}\cdot\frac{a-1}{N-1}+\frac{a}{N}\cdot\frac{b}{N-1}=\frac{a}{N} & \frac{b}{N}\cdot\frac{a}{N-1}+\frac{b}{N}\cdot\frac{b-1}{N-1}=\frac{b}{N}\end{vmatrix}$$

$$X_2\begin{vmatrix}1 & 0\\ \frac{a}{N}\cdot\frac{a-1}{N-1}+\frac{b}{N}\cdot\frac{a}{N-1}=\frac{a}{N} & \frac{a}{N}\cdot\frac{b}{N-1}+\frac{b}{N}\cdot\frac{b-1}{N-1}=\frac{b}{N}\end{vmatrix}$$

Isto quer dizer que, dado um indivíduo u ε U, a probabilidade dele pertencer a uma amostra de grandeza n é sempre igual, seja a amostra obtida com ou sem reposição. A probabilidade de obter uma determinada amostra é, obviamente, diferente, consequência das diferentes distribuições conjuntas.

Exemplo 2 - Consideremos agora uma urna com N bolas, numeradas de 1 a N. Pretendemos calcular as distribuições conjuntas correspondentes a uma amostra de n bolas, sendo n muito menor do que N (n << N). Pretende-se calcular as probabilidades de obter, com e sem reposição, a mesma sequência de n bolas.



Com reposição

$$\frac{1}{N}\ \frac{1}{N}\ \frac{1}{N}\ \ldots\frac{1}{N} = \frac{1}{N^n} = \frac{1}{\overline{\overline{A_n}}} \qquad (1)$$

Sem reposição

$$\frac{1}{N}\ \frac{1}{N-1}\ \frac{1}{N-2}\ \ldots\frac{1}{N-n+1} = \frac{1}{A_n^N} \qquad (2)$$

Se n << N, os valores (1) e (2) são praticamente iguais e podemos portanto utilizar indiscriminadamente amostras obtidas com ou sem reposição. Em populações de dimensão pequena (N pequeno), contudo, já é preciso ter cautela com as correcções para o caracter finito da amostra (ver 1.7.1.2)

2.3.2 Amostragem selectiva

Numa amostragem selectiva a selecção dos indivíduos é feita por qualquer processo que inclui subjectividade na escolha, impossibilidade de selecção de alguns dos indivíduos da população ou qualquer outro mecanismo que impossibilite a atribuição de probabilidades de selecção conhecidas a cada amostra possível. Como consequência, não há possibilidade de calcular um intervalo de confiança para as estimativas feitas e, portanto, não podemos calcular os erros destas estimativas. Este facto não quer dizer que a amostragem selectiva seja de por totalmente de lado. Podem citar-se exemplos em que este processo de selecção é o mais adequado. Suponha-se que se pretende determinar a presença ou ausência de uma determinada praga ou doença numa região. Para tal optou-se por cartografar todos os povoamentos da espécie em questão e dividi-los em parcelas de 100 x 100 m (1 ha). Se a praga ou doença existir na região, mas com uma fraca intensidade (ou seja, com poucos povoamentos parcialmente atacados), a selecção aleatória de parcelas para inspecção pode levar-nos a concluir que a doença não ocorre na região. Pelo contrário, se seleccionarmos para análise, em cada povoamento, as parcelas em que se verificarem árvores secas ou com pouca vitalidade, teremos maior probabilidade de fazer uma avaliação correcta. A amostragem selectiva é utilizada apenas em alguns casos particulares, pelo que não será tratada nestes apontamentos.

2.3.3 Amostragem sistemática

A amostragem sistemática é um método em que a selecção dos indivíduos para análise se baseia numa regra pré-definida - por exemplo, 1 indivíduo em cada 5 – de tal modo que, determinado o primeiro indivíduo da amostra, todos os outros ficam conhecidos. O primeiro indivíduo tem que ser seleccionado por um processo aleatório. A amostragem sistemática é de grande utilidade no inventário florestal, pelo que será objecto do capítulo 10.

No inventário florestal, principalmente por razões de ordem prática, é usual utilizar métodos de selecção da amostra do tipo sistemático. Num método de selecção sistemático as árvores ou parcelas que fazem parte da amostra são seleccionadas de acordo com um padrão ou regra previamente definido, em vez de sorteados com base num processo aleatório.

Vejamos um exemplo. Suponha-se que os N indivíduos da população estão numerados de 1 a N. Para obter uma amostra de n indivíduos, escolha-se um número aleatório a entre 1 e k ($k = (N/n) + 1$), ficando a amostra constituída pelos indivíduos a, $a+k$, $a+2k$, etc. Se N é múltiplo de n, então $k = N/n$. A amostra fica totalmente dependente do primeiro indivíduo sorteado. Este

método de amostragem, designado por amostragem sistemática, não é uma amostragem totalmente selectiva, desde que o primeiro indivíduo seja obtido casualmente. Uma vez que *N* não é geralmente um múltiplo de *n*, as *k* amostras sistemáticas que se podem obter a partir da mesma população finita podem diferir de dimensão a menos de uma unidade. Por exemplo, se *N*=15 e *n*= 4 então $k = \text{int}(15/4) + 1$=4. Temos então 4 amostras de dimensão 4, 4, 4 e 3: (1, 5, 9, 13), (2, 6, 10, 14), (3, 7, 11, 15), (4, 8, 12). Os problemas resultantes deste facto são no entanto desprezáveis desde que *n* > 50 (Cochran, 1977).

2.3.3.1 Vantagens e desvantagens em relação com a amostragem casual

Do ponto de vista prático, a amostragem sistemática tem algumas vantagens sobre a amostragem casual simples:

- Uma vez que a selecção dos indivíduos que fazem parte da amostra é feita com base num padrão ou regra previamente estabelecido, não é necessário identificar todos os indivíduos da população, como é o caso na amostragem casual simples para se poder realizar o sorteio.
- É geralmente mais fácil localizar os indivíduos que fazem parte da amostra.
- Intuitivamente, a amostragem sistemática parece ser mais precisa do que a amostragem casual simples, uma vez que estratifica a população em *n* estratos (com *k* unidades cada), amostrando um indivíduo em cada estrato. De facto, para uma igual grandeza da amostra a amostragem sistemática garante uma melhor "cobertura" da população, evitando a acumulação de indivíduos amostrados numa determinada zona da população, enquanto que outras zonas ficam muito mal representadas na amostra. este último problema pode ocorrer com alguma frequência na amostragem casual simples, especialmente no caso da dimensão da amostra ser pequena em relação com a dimensão da população.

Há, contudo, um problema associado com a amostragem sistemática. Uma vez que os indivíduos não são seleccionados aleatóriamente, não é possível, do ponto de vista teórico, estimar o intervalo de confiança para a média. Esta dificuldade só pode ser ultrapassada em populações nas quais se possa admitir que os indivíduos se encontram aleatoriamente distribuido na população, o que sabemos não ser completamente verdade. Existe sempre alguma tendência para os valores do atributo de indivíduos (parcelas) vizinhos estarem correlacionados. Scheffar et al. (1990) mostraram, em diversos tipos de populações, que os estimadores da variância da média para a amostragem casual podem ser utilizados sem problema quando a amostra é obtida por um

processo sistemático. Esta conclusão não é válida para as populações caracterizadas por algum tipo de variação periódica. Neste caso, os estimadores da variância da média para a amostragem casual são geralmente enviesados por defeito. Há, portanto, que analisar a possibilidade de ocorrência de variações sistemáticas numa população, antes de decidir qual o processo de selecção da amostra a utilizar.

2.3.3.2 Implementação prática de uma amostragem sistemática em inventário florestal

Do ponto de vista operacional as amostragens sistemáticas são geralmente planeadas de acordo com uma malha (grelha) quadrada ou rectangular. A forma da malha depende essencialmente das características da topografia da zona a amostrar. Em zonas planas prefere-se uma malha quadrada, em zonas com um declive acentuado, prefere-se uma malha rectangular, com o comprimento dos rectângulos perpendicular à encosta (para diminuir as deslocações no sentido do maior declive).

Suponhamos que queremos seleccionar uma amostra de 48 parcelas na população I –talhão de uma Mata Nacional. A área correspondente a cada parcela será de:

$$area = \frac{Area\ total}{48} = \frac{40}{48} = 0.8333\ \ ha = 8333.33\ \ m^2$$

Uma vez que a área de um quadrado é igual ao quadrado do lado, o lado de uma malha será então:

$$lado = \sqrt{8333.33} = 91.28\ \ m \approx 90\ \ m$$

Há então que determinar aleatoriamente a localização da primeira parcela e, em seguida, basta andar, paralelamente aos lados do talhão (na prática de acordo com um azimute), 90 m e medir uma nova parcela. Ao chegar aso limites do talhão, deslocamo-nos 90 m numa direcção perpendicular à primeira para encontrarmos a parcela de partida para a segunda linha de parcelas. A localização aleatória da localização da primeira parcela pode ser feita, por exemplo, do seguinte modo (figura 11):

1. sorteia-se um número aleatório entre 0 e 90 m. Este número indica a distância desde um dos cantos do talhão até à primeira linha de parcelas (dist 1)

2. sorteia-se um segundo número aleatório entre 0 e 90 m. este número indica a distância, na primeira linha de parcelas, desde o limite do talhão até ao local da 1ª parcela (dist 2)

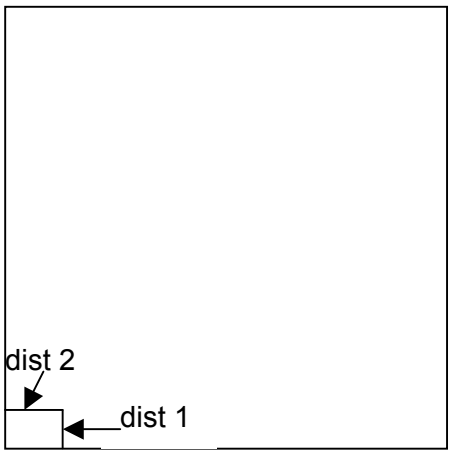


Figura 10. Localização aleatória da primeira parcela de uma amostragem sistemática

O planeamento de uma malha rectangular é bastante semelhante, baseando-se geralmente na fixação prévia da relação entre os lados do rectângulo. Por exemplo, pode seleccionar-se que um dos lados é 4 vezes maior que o outro. Assim, e desingando o lado menor por *ladom* e o lado maior por *ladoM*, temos que:

$$area = ladom * 4\ \ ladom = 4\ \ ladom^2$$

$$ladom = \sqrt{\frac{8333.33}{4}} = 45.64 \approx 45\ \ m$$

$$ladoM = 4\ \ ladom = 180\ \ m$$

A selecção do local para a primeira parcela pode fazer-se por um processo semelhante ao sugerido para a malha quadrada.

2.4 Função de distribuição empírica da amostra

Seja $(x_1,...,x_n)$ uma amostra de grandeza n da v.a. X.

Podemos sempre passar para a amostra ordenada correspondente *(z_1, z_2,...,z_n)*, onde

$$\left.\begin{array}{l} z_1 = \min(x_1,...,x_n) \\ z_2 = \min(x_1,...,x_n) \\ ... \\ z_n = \min(x_1,...,x_n) \end{array}\right] \qquad \text{Estatísticas ordinais}$$

Com base nas estatísticas ordinais determina-se facilmente a função de distribuição empírica da amostra

$$S_n(X) = \frac{k_x}{n} \qquad k_x \text{ – n.º de valores observados} \leq x$$

A figura 11 representa a função de distribuição empírica da amostra de 48 indivíduos obtida a partir da população I descrita no ponto 1.2 .

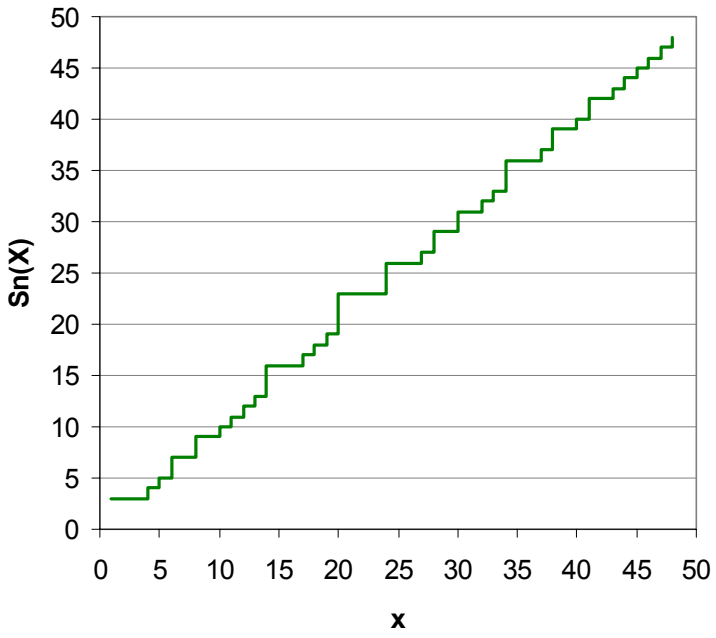


Figura 11. Função de distribuição empírica de uma amostra de 48 indivíduos da população I

No fundo $S_n(x)$ representa a frequência relativa de se observarem valores $\leq x$

$$S_n(X) = \frac{k_x}{n} = f_r(X \leq x)$$

Por definição, a função de distribuição da variável *X* é, como sabemos

$$F(x) = P(X \leq x)$$

Mas o teorema de Bernoulli diz-nos que

$$f_r(X \leq x) \xrightarrow[n \to \infty]{P} P(X \leq x)$$

Logo:

$$S_n(x) = \frac{k_x}{n} \xrightarrow{P} F(x)$$

Isto significa que, desde que a dimensão da amostra seja razoável, os dois gráficos tendem a confundir-se.

Diz-se então que $S_n(x)$ dá uma imagem estatística de $F(x)$, ou seja, a função de distribuição empírica da amostra dá uma imagem estatística da função de distribuição da variável aleatória *X*.

2.5 Etapas para a realização prática de uma amostragem

Para a realização prática de uma amostragem é necessário ter em atenção vários aspectos fundamentais, uns de preparação, outros de execução. Assim, vejamos as principais etapas para a sua realização:

- Definição de objectivos
- Definição da população
- Definição do esquema de amostragem
- Definição do método de selecção da amostra
- Cálculo da grandeza da amostra
- Selecção dos indivíduos que fazem parte da amostra
- Análise dos indivíduos da amostra
- Tratamento de dados

Definição de objectivos:

Antes de se iniciar qualquer trabalho de amostragem há que definir claramente qual a variável (ou variáveis) que se pretende avaliar, com que erro e para que nível de significância. O objectivo de uma amostragem pode ser a média ou o total de uma variável quantitativa, a proporção de indivíduos com uma determinada característica, etc.

Definição da população:

Como já vimos, população é o universo do qual a amostra vai ser escolhida. Já vimos também que há casos em que a definição da população não levanta qualquer problema enquanto que noutros há que definir não só o indivíduo, como regras que permitam decidir em cada caso se um determinado indivíduo pertence ou não à população. Esta é uma das etapas de planeamento de uma amostragem de maior importância nas amostragens realizadas no âmbito de um inventário florestal. A correcta definição dos indivíduos pode ter uma influência bastante grande na variabilidade da população, a qual é essencial para o erro de amostragem que se pode obter. Por exemplo, na avaliação de volume de uma mata, se optarmos por realizar umamostragem com base em árvores individuais, teremos uma variabilidade entre indivíduos bastante maior do que se a basearmos em parcelas de 500 m^2. Parte da variabilidade entre indivíduos é "absorvida" ao juntarmos vários indivíduos em parcelas. Na maior parte dos povoamentos, conseguiremos ainda uma diminuição da variabilidade se utilizarmos parcelas de 1000 m^2. Contudo, podemos não conseguir diminuir grandemente a variabilidade entre indivíduos se aumentarmos a dimensão das parcelas para 2000 m^2. A dimensão óptima das parcelas de amostragem depende, óbviamente, das características do povoamento em estudo.

Definição do esquema de amostragem:

Já vimos que os resultados de uma amostragem se expressam em termos de um intervalo de confiança, cujo erro depende da dimensão da amostra e da variabilidade da população. Já vimos também que temos alguma liberdade na definição dos indivíduos que constituem a população a amostrar, o que nos permite ter algum controlo sobre a variabilidade da população. De acordo com características de cada população podemos ainda recorrer a esquemas de amostragem que

permitem obter reduções da variabilidade entre os indivíduos da população, permitindo assim a diminuição do erro de amostragem para o mesmo número de indivíduos analisados. Nos capítulos seguintes (5 a 9) descrevem-se alguns dos esquemas de amostragem mais utilizados em inventário florestal.

Definição do método de selecção dos indivíduos da amostra:

Como já vimos, os indivíduos a amostrar podem ser seleccionados por diferentes processos: amostragem aleatória, amostragem selectiva ou amostragem sistemática. É nesta fase que se deve optar por um destes métodos de selecção.

Cálculo da dimensão da amostra:

Antes de se iniciar uma amostragem, é importante tentar determinar qual a dimensão (*n*) que a amostra deve ter para permitir atingir um erro próximo daquele que se pretende. Sendo, como vimos, o resultado de uma amostragem expresso sob a forma de um intervalo de confiança, no qual o erro E é função da grandeza da amostra (*n*) é possível determinar aproximadamente a grandeza da amostra que garanta um erro percentual próximo do pretendido. Nos capítulos seguintes (6 a 10) veremos, para cada um dos esquemas de amostragem estudados, qual o correspondente método para o cálculo da grandeza da amostra.

Selecção dos indivíduos que fazem parte da amostra:

Uma vez determinado o número de indivíduos que fazem parte da amostra, há que seleccioná-los de acordo com o método de selecção previamente seleccionado: amostragem aleatória, selectiva ou sistemática. Muitas vezes, a implementação prática de um método de selecção dos indivíduos que fazem parte da amostra não é fácil, especialmente se a população a estudar for bastante grande. Ao longo do texto, assim como nos exercícios práticos que acompanham estes apontamentos, procuraremos apontar algumas das soluções mais utilizadas para a selecção dos indivíduos que fazem parte da amostra, no caso da amostragem aleatória e da amostragem sistemática.

Avaliação da variável em cada um dos indivíduos da amostra:

Uma vez seleccionados os indivíduos que fazem parte da amostra, há que obter as medições de acordo com métodos previamente definidos, no caso do inventário florestal de acordo com os diversos métodos de medição de árvores e povoamentos disponíveis.

Tratamento de dados:

Finalmente, há que realizar o tratamento de dados cujo resultado final, como já vimos, é um intervalo de confiança para a média da variável que se pretende avaliar.

2.6 Exercícios

2.6.1 Amostragem aleatória

Ex2.1 Obtenha uma amostra de 48 indivíduos da população I:

a) Sem reposição

b) Com reposição

c) Calcule a média empírica da amostra para cada uma das amostras obtidas

Ex2.2 Faça a distribuição empírica da amostra para cada uma das amostras obtidas no exercício anterior e compare-a com a distribuição acumulada da população que obteve no exercício **Ex1.7**.

2.6.2 Amostragem sistemática

Ex2.3 Esquemas de obtenção de amostras sistemáticas

a) Sugira alguns esquemas alternativos para a obtenção de amostras sistemáticas da população I, com aproximadamente 48 indivíduos.

b) Calcule a média empírica da amostra para cada uma das amostras que obteve

Ex2.4 Idêntico a 1, mas em relação à população 2. Procure interpretar as diferenças encontradas.

3 Teoria da estimação

3.1 Estatísticas

3.1.1 Definição

Consideremos uma população na qual estamos interessados através de uma característica mensurável ou variável aleatória *X* cuja distribuição é *F(x).*

Seja $(x_1,...,x_n)$ uma amostra de dimensão n, valor observado de um vector aleatório $\vec{X}$ de $\vec{n}$ componentes independentes e semelhantes. Uma estatística é uma função dos valores observados que não dependa de quaisquer parâmetros desconhecidos.

Uma estatística é então uma variável aleatória, uma vez que é uma função de variáveis aleatórias. Utiliza-se uma letra maiúscula para designar uma estatística, reservando-se a letra minúscula correspondente para designar um valor determinado, correspondente a uma determinada amostra.

$$Z = \varphi(X_1, X_2, ..., X_n) = \varphi\left(\vec{X}\right) \qquad \text{estatística Z}$$

$$z = \varphi(x_1, x_2, ..., x_n) = \varphi\left(\vec{x}\right) \qquad \text{valor que Z toma para a amostra } \vec{x}$$

3.1.2 Distribuições de amostragem

A distribuição de amostragem é a função de distribuição de uma estatística, definida sobre a população (finita ou infinita) das amostras de dimensão *n*. Esta distribuição depende essencialmente de:

- *F(x),* função de distribuição da população

- *x*, função que define a estatística

Muitas vezes não se consegue determinar a expressão exacta da distribuição de amostragem duma estatística (ou é de tal modo complicada que perde o interesse prático), mas consegue

determinar-se a distribuição assimptótica de amostragem, ou seja, a distribuição limite quando $\underline{n}$ (grandeza da amostra) tende para infinito:

Seja $Z(\overline{x_k})$ a estatística para uma amostra de dimensão k e G_k a distribuição correspondente. Obtém-se então uma sucessão de distribuições

$$G_1, G_2, G_3, ..., G_n \quad \underset{n\to\infty}{\rightarrow} \quad G$$

G, limite da sucessão $\{G_k\}$ quando $n\to\infty$ é a distribuição assimptótica de amostragem de Z.

3.1.3 Estatísticas correspondentes aos parâmetros

A cada parâmetro da população (média, variância, mediana, percentis, etc.) corresponde uma estatística, calculada com os valores da amostra por processo idêntico ao do cálculo dos parâmetros; esta recebe o nome do parâmetro correspondente acrescido do adjectivo empírico. A tabela 4 apresenta as estatísticas mais importantes para a teoria da amostragem.

Grande número dos parâmetros duma variável aparecem como valores médios da própria variável ou duma função da variável. Prova-se que a distribuição das estatísticas correspondentes a estes parâmetros é assimptoticamente normal.

3.2 Estimação pontual e por intervalos

Consideremos uma população X com densidade $f(x;\theta)$ conhecida, mas dependente de um parâmetro θ desconhecido (a generalização para mais de um parâmetro desconhecido é simples).

A estimação é a avaliação, com base na observação de uma amostra, do valor real do parâmetro θ desconhecido.

Há dois tipos de estimação:

- Estimação pontual - procura-se directamente o valor do parâmetro, através dum estimador.
- Estimação regional ou por intervalos - procura-se um intervalo que contenha o verdadeiro valor do parâmetro, com um probabilidade conhecida.

Tabela 4. Estatísticas correspondentes aos parâmetros mais importantes de uma população

Parâmetro[1]	Estatística correspondente
Média $\mu_1' = \mu = E(X) = \frac{1}{N}\sum_{i=1}^{N} x_i$	**Média empírica** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Momento de ordem k $\mu_k' = E(X^k) = \frac{1}{N}\sum_{i=1}^{N} x_i^k$	**Momento empírico de ordem k** $m_k' = \frac{1}{n}\sum_{i=1}^{n} x_i^k$
Variância $\mu_2 = \sigma^2 = E(X-\mu)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$	**Variância empírica** $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_1-\bar{x}.)^2$
Momento central de ordem k $\mu_2^k = E(X-\mu)^k = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^k$	**Momento central empírico de ordem k** $m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k$

[1] As expressões sob a forma de somatórios correspondem a populações finitas

3.2.1 Estimação pontual

O objectivo da estimação pontual é, como foi dito, encontrar, com base nos valores observados nos indivíduos que fazem parte da amostra, um valor que esteja próximo do parâmetro θ. Parece lógico que esta avaliação seja conseguida com uma função dos valores observados na amostra que não dependa de θ, ou seja, de uma estatística. Um estimador não é, assim, mais do que uma estatística convenientemente escolhida Θ^*.

$$\Theta^* = \Theta = (X_1,...,X_n) = \Theta^*\left(\vec{X}\right)$$

A um valor bem determinado de Θ^*, correspondente a uma determinada amostra, chama-se uma estimativa θ^*.

$$\theta^* = \theta^*(x_1, ..., x_n)$$

Na prática utiliza-se frequentemente a letra minúscula θ^* para ambos os conceitos - estimador e estimativa - mas é importante fixar a diferença que existe entre eles.

Vejamos o que se entende por uma estatística convenientemente escolhida. Uma estatística convenientemente escolhida deve gozar das duas propriedades seguintes: convergência ou consistência e suficiência. Além disso, a distribuição de amostragem de θ^* deve ser tão concentrada quanto possível em torno do verdadeiro valor real do parâmetro θ. Quer isto dizer que se calculassemos sucessivas estimativas de θ com base em sucessivas amostras, grande número destas estimativas deveria estar próximo do valor real do parâmetro.

3.2.1.1 Convergência ou consistência

A propriedade da convergência implica que, à medida que a dimensão da amostra aumenta, a probabilidade de se obterem estimativas mais próximas do valor do parâmetro também aumenta.

Seja $\theta^*_{(k)}$ a estatística para uma amostra de dimensão k. Temos então a sucessão:

$$\theta^*_{(1)}, \theta^*_{(2)},..., \theta^*_{(n)}...$$

De modo formal, a propriedade da convergência expressa-se por

$\left\{\theta_n^*\right\} \xrightarrow{p} \theta$ a sucessão $\left\{\theta_n^*\right\}$ converge em probabilidade para θ.

ou seja, $P\left(\left|\theta_n^* - \theta\right| \langle \varepsilon\right) \underset{n\to\infty}{\longrightarrow} 1 \qquad \forall\ \varepsilon \rangle 0$

3.2.1.2 Suficiência

O cálculo de uma estatística implica sempre alguma perca de informação. Uma estatística suficiente esgota toda a informação contida na amostra sobre o parâmetro a estimar θ. Em termos formais, pode dizer-se que uma estatística Θ^* é suficiente para θ se e só se, dada qualquer outra estatística se verifica a seguinte condição:

$\varphi\left(\hat{\theta} / \theta^*\right)$ é independente de θ, $\forall\ \hat{\theta}$

Esta definição é, de facto, lógica pois implica que uma vez calculado θ^* se esgota toda a informação contida na amostra sobre o parâmetro θ. $\hat{\theta}$ já não acrescenta qualquer outra informação.

3.2.1.3 Propriedades dos estimadores

Seja $g(\theta^*,\theta)$ a densidade de probabilidade de θ^*. É dependente de θ, visto o mesmo se passar com $f(x;.\ \theta)$

Intuitivamente, devemos esperar que a distribuição de amostragem de θ^* esteja concentrada em torno de θ. Esta exigência traduz-se pela condição:

$E\ (\theta^*-\theta)^2=\min$ erro quadrado médio mínimo

Analisemos então $E\ (\theta^*-\theta)^2$:

$$E\left(\theta^* - \theta\right)^2 = E\left[\left(\theta^* - E\left(\theta^*\right)\right) + \left(E\left(\theta^*\right) - \theta\right)\right]^2$$

$$= \underbrace{E\left(\theta^* - E\left(\theta^*\right)\right)^2}_{\text{variância de } \theta^*} + \underbrace{E\left[E\left(\theta^*\right) - \theta\right]^2}_{\substack{\text{valor médio} \\ \text{do quadrado} \\ \text{do desvio}}}$$

Vemos assim que o erro quadrado médio tem duas componentes não negativas, a variância de θ^* e o valor médio do quadrado do seu desvio em relação ao parâmetro a estimar θ. A primeira componente expressa a variabilidade dos desvios, indicando um grande valor que o estimador tem uma elevada probabilidade de originar estimativas muito afastadas do verdadeiro valor de θ. A segunda componente, por seu lado, expressa a tendência (se diferente de zero) para o estimador originar estimativas por excesso ou por defeito.

A figura 12 ilustra as duas componentes do erro quadrático médio, através de dois exemplos. Em (A) representam-se as densidades de probabilidade de dois estimadores com diferentes variâncias, sendo a segunda componente do erro quadrático médio nula em ambos os casos. Em (B) representam-se as densidades de probabilidade de dois estimadores com diferentes variâncias, apresentando o estimador com menor variância um pequeno desvio. Em (A) não há dúvida de que o estimador com menor variância deve ser preferido, em (B) a decisão dependerá da grandeza do desvio.

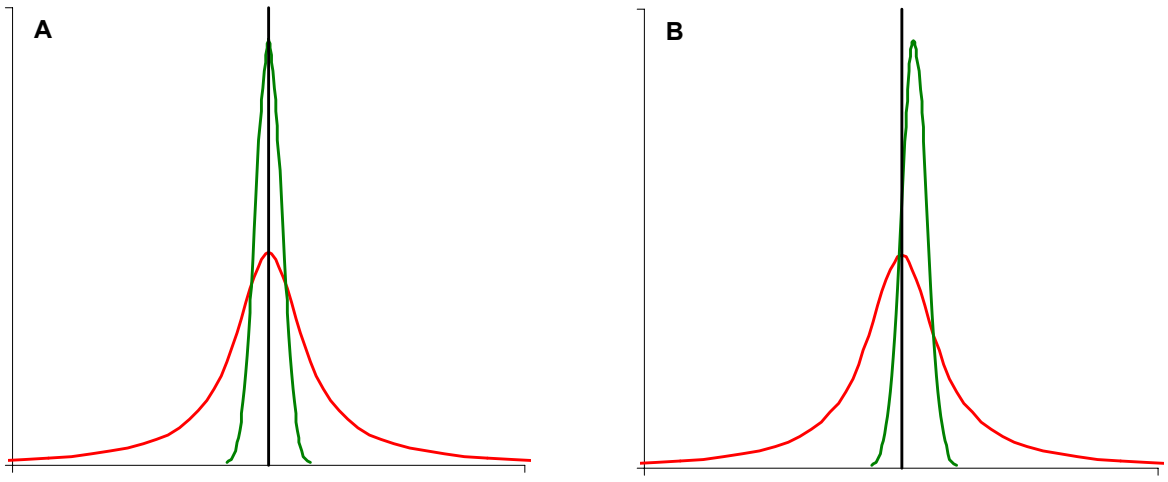


Figura 12. Exemplos de densidades de probabilidade de estimadores com diferentes valores de variância e de desvio (ver texto).

Um estimador deve, assim, gozar de três propriedades:

1. Um estimador deve ser centrado ou, pelo menos, assimptoticamente centrado.

   Um estimador será centrado se se anular a segunda componente, isto é, se o desvio for nulo. Neste caso o erro quadrado médio coincide com a variância.

$$E(\theta^*) - \theta = 0 \quad \Leftrightarrow \quad E(\theta^*) = \theta$$

Um estimador é assimptoticamente centrado se

$$\lim_{n \to \infty} \left(E(\theta^*) - \theta \right) = 0$$

2. Um estimador deve gozar da propriedade da variância mínima.

   Esta propriedade minimiza a primeira componente do erro quadrado médio. Um estimador de variância mínima é geralmente preferido a um estimador centrado, desde que o desvio não seja muito grande.

3. Um estimador deve ser eficiente relativamente a outros ou, pelo menos, assimptoticamente eficiente.

   Dados dois estimadores θ_1^* e θ_2^* ambos calculados sobre amostras de dimensão *n*, define-se eficiência de θ_2^* relativamente a θ_1^* como:

$$ef = \frac{E\left(\theta_2^* - \theta\right)^2}{E\left(\theta_1^* - \theta\right)^2}$$

   Uma eficiência inferior a 1, indica que θ_2^* é mais eficiente que θ_1^*.

   Define-se também a eficiência assimptótica de $\theta_{(n)}^*$ relativamente a $\theta_{(n)}^1$:

$$ef_\infty = \lim_{n\to\infty} \frac{E\left(\theta_{(n)}^* - \theta\right)^2}{E\left(\theta_{(n)}^1 - \theta\right)^2}$$

   Pode acontecer que até determinado valor n_0, θ_1 seja uma melhor estimativa, invertendo-se depois os papeis para $n>n_0$. Diz-se que um estimador θ^* é assimptóticamente eficiente, se:

$$\lim_{n\to\infty} \frac{E\left(\theta_{(n)}^* - \theta\right)^2}{E\left(\theta_{(n)}^1 - \theta\right)^2} \leq 1 \qquad \forall\ \theta_{(n)}^1$$

3.2.1.4 Estimadores para a média, variância e variância da média (amostragem com reposição)

Prova-se facilmente que a média da amostra ($\overline{X}$) é um estimador centrado de μ:

$$\begin{aligned} E\left(\overline{X}\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \frac{1}{n} n\ \mu = \mu \end{aligned}$$

Comecemos por calcular uma expressão para a variância da variável aleatória média da amostra:

$$
\begin{aligned}
var\left(\overline{X}\right) &= \sigma_{\overline{X}}^{2} \\
&= var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2} var\left(\sum_{i=1}^{n} X_i\right) && X_i \text{ independentes} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} var\left(X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\
&= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}
\end{aligned}
$$

Pode agora calcular-se o valor médio da variância da amostra (s^2):

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

$$
\begin{aligned}
E\left(s^2\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - E\left(\overline{X}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^2\right) - E\left(\overline{X}^2\right) = (A)
\end{aligned}
$$

Servindo-nos agora da definição de variância:

$$
\begin{aligned}
&var\left(X_i\right) = E\left(X_i - E\left(X_i\right)\right)^2 = E\left(X_i^2\right) - E^2\left(X_i\right) \\
&\Downarrow \\
&E\left(X_i^2\right) = Var\left(X_i\right) + E^2\left(X_i\right)
\end{aligned}
$$

Temos que:

$$
\begin{aligned}
(A) &= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 + \mu^2 - \left(var\left(\overline{X}\right) + \mu^2\right) \\
&= \sigma^2 - \frac{\sigma^2}{n} \\
&\sigma^2\,\frac{n-1}{n}
\end{aligned}
$$

Conclui-se assim que s^2 não é um estimador centrado de σ^2, embora seja assimptoticamente centrado:

$$\lim_{n\to\infty} E\left(s^2\right) = \lim_{n\to\infty} E\left(\sigma^2\left(\frac{n-1}{n}\right)\right) = \sigma^2$$

Atendendo ao facto de que $E(s^2) = \sigma^2(n-1)/n$ pode deduzir-se facilmente o seguinte estimador centrado de σ^2:

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

$$E\left(s_c^2\right) = E\left(s^2 \frac{n}{n-1}\right) = \frac{n}{n-1}E\left(s^2\right) = \frac{n}{n-1}\sigma^2 \frac{n-1}{n} = \sigma^2$$

Com base no estimador centrado da variância s_c^2, temos então o seguinte estimador centrado da variância da média:

$$s_{\overline{X}}^2 = \frac{s_c^2}{n}$$

$$E\left(s_{\overline{X}}^2\right) = E\left(\frac{s_c^2}{n}\right) = \frac{1}{n}E\left(s_c^2\right) = \frac{\sigma^2}{n}$$

3.2.1.5 Factor de correcção para os estimadores da média, variância e variância da média quando a amostragem é feita sem reposição

Seja *U* uma população finita com *N* indivíduos e *X* uma característica mensurável dos indivíduos de *U*, tomando os seguintes valores y_1, y_2, ..., y_n.

Seja (x_1, x_2, ..., x_n) uma amostra obtida sem reposição.

Defina-se uma variável aleatória t_i, variável indicatriz do acontecimento “escolha de um indivíduo”:

$$t_i = \begin{cases} = 1 & \text{se o indivíduo faz parte da amostra} \\ = 0 & \text{se o indivíduo não faz parte da amostra} \end{cases}$$

Podemos então escrever as seguintes expressões para a média e variância empíricas:

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

$$s^2 = \frac{1}{n}\sum_{i=1}^{N} t_i (y_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{N} t_i \, y_i{}^2 - \bar{x}^2$$

Vejamos agora qual será a probabilidade do indivíduo *i* fazer parte de uma amostra de *n* elementos obtida sem reposição ($\Rightarrow$ distribuição de probabilidade da v.a. t_i): o número de amostras de dimensão *n* que se podem tirar duma população com *N* indivíduos é igual a $\binom{N}{n}$

O indivíduo *i* entra em $\binom{N-1}{n-1}$ destas amostras, pelo que:

$$P(t_i = 1) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

$$P(t_i = 0) = 1 - P(t_i = 1) = 1 - \frac{n}{N}$$

Tem-se ainda que

$$P(t_i \; t_j = 1) = P(t_i = 1) \; P(t_i = 1 / t_i = 1) = \frac{n}{N} \; \frac{n-1}{N-1}$$

$$E(t_i) = E(t_i{}^2) = \frac{n}{N} \; x \; 1 + \left(1 - \frac{n}{N}\right) \; x \; 0 = \frac{n}{N}$$

$$E(t_i \; t_j) = \frac{n}{N} \; \frac{(n-1)}{(N-1)} \; x \; 1 + \left(1 - \frac{n}{N} \; \frac{(n-1)}{(N-1)}\right) \; x \; 0 = \frac{n}{N} \; \frac{(n-1)}{(N-1)}$$

Calculemos agora $E(\bar{x})$ e $v(\bar{x})$:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{N} t_i y_i\right)$$
$$= \frac{1}{n}\ E\left(\sum_{i=1}^{N} t_i y_i\right)$$
$$= \frac{1}{n}\sum_{i=1}^{N} y_i\ E(t_i)$$
$$= \sum_{i=1}^{N} \frac{y_i}{n}\ \ \frac{n}{N} = \sum_{i=1}^{N} \frac{y_i}{N} = \mu$$

$$V(\bar{X}) = E\left(\frac{1}{N}\sum_{i=1}^{N} t_i y_i - \mu\right)^2$$
$$= E\left(\frac{1}{n}\sum_{i=1}^{N} t_i y_i\right)^2 - \mu\ E(\bar{X}) + \mu^2$$
$$= \frac{1}{n^2}\ E\left(\sum_{i=1}^{N} t_i y_i\right)^2 - \mu^2$$
$$= \frac{1}{n^2}\ E\left(\sum_{i=1}^{N} t_i^2 y_i^2\right) + \frac{1}{n^2}\ E\left(\sum_{\substack{i,j=1 \\ i\neq j}}^{n} t_i\ t_j\ y_i\ y_j\right) - \mu^2$$

$$= \frac{1}{n^2}\ \sum_{i=1}^{N} y_i^2 E(t_i^2) + \frac{1}{n^2}\sum_{\substack{i,j \\ i\neq j}} y_i\ y_j\ \mathrm{E}E(t_i\ t_j) - \mu^2$$
$$= \frac{1}{n^2}\ \frac{n}{N}\sum_{i=1}^{N} y_i^2 + \frac{1}{n^2}\frac{n(n-1)}{N(N-1)}\sum_{\substack{i,j \\ i\neq j}} y_i\ y_j - \mu^2$$
$$= (A)$$

Mas

$$\sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N} (y_i - \mu)^2 + N\mu^2$$
$$= N\ v(x) + N\mu^2 = N(\sigma^2 + \mu^2)$$

$$\sum_{\substack{i,j \\ i\neq j}} y_i y_j = \left(\sum_i y_i\right)^2 - \sum_i y_i^2$$
$$= N^2\mu^2 - N(\sigma^2 + \mu^2)$$

Pelo que

$$
\begin{aligned}
(A) &= \frac{1}{nN} N\left(\sigma^2 + \mu^2\right) + \frac{(n-1)}{nN(N-1)}\left[N^2\mu^2 - N\left(\sigma^2 + \mu^2\right)\right] - \mu^2 \\
&= \frac{1}{n}\left(\sigma^2 + \mu^2\right) + \frac{(n-1)}{n(N-1)}\left[N\mu^2 - \sigma^2 - \mu^2\right] - \mu^2 \\
&= \frac{1}{n}\left(\sigma^2 + \mu^2\right) + \frac{n-1}{n(N-1)}(N-1)\mu^2 - \frac{n-1}{n(N-1)}\sigma^2 - \mu^2 \\
&= \frac{1}{n}\sigma^2 + \frac{1}{n}\mu^2 + \frac{(n-1)\mu^2}{n} - \mu^2 - \frac{n-1}{n(N-1)}\sigma^2 \\
&= \frac{\sigma^2}{n} - \frac{n-1}{n(N-1)}\sigma^2 \\
&= \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) = \frac{\sigma^2}{n}\frac{N-n}{N-1}
\end{aligned}
$$

Quando σ^2 na expressão atrás, não é conhecido $v(\overline{X})$ é estimada pela mesma expressão mas com σ^2 substituído pelo seu estimador

$$s^2 = \frac{1}{n}\sum_{i=1}^{N} t_i \; y_i^2 - \overline{x}^2$$

Vejamos se s^2 é centrado:

$$
\begin{aligned}
E\left(s^2\right) &= E\left(\frac{1}{n}\sum_{i=1}^{N} t_i {y_i}^2 - \overline{X}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{N} {y_i}^2 E(t_i) - E\left(\overline{X}^2\right) \\
&= \frac{1}{n}\ \frac{n}{N}\sum_{i=1}^{N} {y_i}^2 - \left[v\left(\overline{X}\right) + E^2\left(\overline{X}\right)\right] \\
&= \frac{1}{N}\ N\left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n}\ \frac{N-n}{N-1} + \mu^2\right) \\
&= \sigma^2 + \mu^2 - \frac{\sigma^2}{n}\ \frac{N-n}{n-1} - \mu^2 \\
&= \sigma^2\ \left(1 - \frac{N-n}{n(N-1)}\right) \\
&= \frac{N(n-1)}{n(N-1)}\ \sigma^2 \\
&= \frac{N}{N-1}\ \frac{n-1}{n}\ \sigma^2
\end{aligned}
$$

Daqui se conclui que

$\frac{N-1}{N}.\frac{n}{n-1}s^2$ é um estimador centrado de $v\left(\overline{X}\right)$

Assim, as expressões a utilizar para $\overline{x}$, s_c^2 e $s_{\overline{x}}^2$ no caso da amostragem sem reposição são:

$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (não se altera)

$$
s_c^2 = \frac{N-1}{N}\ \frac{n}{n-1}s^2 = \frac{N-1}{N}\ \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2
$$

$$s_{\bar{x}}^2 = \frac{N-1}{N} \frac{n}{n-1} s^2 x \frac{1}{n} \frac{N-n}{N-1}$$

$$= \frac{s_c^2}{n} \frac{N-n}{N}$$

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

$f = \frac{n}{N}$ designa-se por fracção de amostragem.

O factor (1-*f*), factor de correcção para a amostragem sem reposição, tem pouco significado quando *f* é muito pequeno, ou seja quando a dimensão da amostra é pequena relativamente à dimensão da população. Na prática, e quando o cálculo seja manual, pode deixar de se utilizar sempre que *f*<0.1.

3.2.2 Estimação por intervalos

A estimação por intervalos faz-se com base na distribuição de amostragem de θ^*.

Vejamos com um exemplo:

Seja $X \cap N\ (\mu, \sigma^2)$, σ^2 conhecido

Já vimos que $\bar{X}$ é um bom estimador de μ

Já vimos também que $\bar{X} \cap N\ (\mu, \sigma^2/n)$, pelo que se pode construir a média reduzida:

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \qquad \cap \qquad N\ (0,1)$$

É então fácil encontrar um intervalo no qual Z esteja com uma determinada probabilidade, geralmente designada por (1-α). Dada a simetria da lei normal este intervalo é do tipo

$$P\left(-z_{\alpha/2} < \sqrt{n}\frac{\bar{x}-\mu}{\sigma} < z_{\alpha/2}\right) = 1-\alpha$$

Daqui deduz-se imediatamente o intervalo para μ, com uma probabilidade correspondente a (1-α):

$$P\left(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} < \mu < \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = 1 - \alpha$$

Este intervalo denomina-se intervalo de confiança, a probabilidade (1-α) é o coeficiente de confiança do intervalo e α designa-se por nível de significância. Os limites do intervalo

$$\bar{x} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

são os limites de confiança.

O método consiste em encontrar uma função Z da amostra e do parâmetro (geralmente uma estatística) cuja distribuição seja independente de quaisquer parâmetros desconhecidos. O intervalo resulta então de

$$P(f_{\alpha_1} < Z < f_{\alpha_2}) = 1 - \alpha \qquad \alpha_1 + \alpha_2 = \alpha$$

Deste intervalo deduz-se então o intervalo para o parâmetro, por manipulação algébrica.

3.2.3 A amostragem e a teoria da estimação

O objectivo de uma amostragem é avaliar o valor médio por indivíduo de uma determinada característica mensurável (basta depois multiplicar pelo número total de indivíduos para obter uma avaliação para o total). A amostragem reduz-se assim a um problema de estimação do valor médio da variável aleatória associada à característica mensurável em questão. Esta estimação é feita através de um intervalo de confiança para a média:

$$P\left(\overline{X}_n + E \leq \mu \leq \overline{X}_n - E\right) = 1 - \alpha$$

A quantidade E costuma geralmente designar-se por erro de amostragem ou erro absoluto. Se este erro for expresso em percentagem de estimativa de valor médio, temos o erro percentual, aquele que é geralmente utilizado para apresentar os resultados de uma amostragem, uma vez que é independente da dimansão da população:

$$E\% = \frac{E}{\overline{X}_n} \; 100$$

O erro de amostragem depende de dois factores:

- a grandeza da amostra *n* (diminuindo se *n* aumenta)
- a variabilidade da população (aumentando com a variabilidade da população)

A construção do intervalo de confiança pode ser mais ou menos complicada, consoante se conhece ou não a função de distribuição da variável aleatória e de acordo com o esquema de amostragem que em cada caso melhor se adapte à população em estudo.

Analisaremos mais à frente os esquemas de amostragem mais utilizados em inventário florestal e os intervalos de confiança respectivos.

Os estimadores da média (μ), variância (σ^2) e variância da média ($\sigma^2_{\bar{x}}$) figuram em todos os intervalos de confiança para a média, pelo que convém rever as suas expressões tanto no caso da amostragem com reposição como no da amostragem sem reposição (tabela 5).

Tabela 5. Estimadores da média, variância e variância da média nas amostragens som e sem reposição

	Amostragem com reposição	**Amostragem sem reposição**
Estimador centrado de μ	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
Estimador centrado de σ^2	$s_c^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$	$s_{s,c}^2 = s_c^2 \ \frac{N-1}{N}$
Variância de $\overline{X}$ - $\sigma^2_{\overline{X}}$	$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$	$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}\frac{N-n}{N-1}$
Estimador centrado de $\sigma^2_{\overline{X}}$	$s^2_{\overline{X}} = \frac{s_c^2}{n}$	$s^2_{\overline{X}} = \frac{s_c^2}{n} \ \frac{N-n}{N}$

3.3 Exercícios

Ex3.1 Estimador da média

Verifique que $\overline{X}$ (média empírica da amostra) é um estimador centrado de μ.

Ex3.2 Estimadores da variância

a) Verifique que s^2 não é um estimador centrado de σ^2, embora seja assimptóticamente centrado.

b) Com base nos resultados da alínea anterior introduza uma correcção em s^2 de modo a transformá-lo num estimador s_c^2 centrado de σ^2.

Ex3.3 O mesmo que **Ex3.1** e **Ex 3.2**, mas em relação à proporção p^* e s_p^2

Nota: $s_p^2 = p^* q^*$

Ex3.4 Distribuição de amostragem da média empírica

a) Verifique que a distribuição de amostragem da média empírica $\overline{X}$ é assimptóticamente normal.

Sugestão: Utilize o teorema limite central

b) Deduza os parâmetros da distribuição de $\overline{X}$, $E(\overline{X})$ e $var(\overline{X}) = \sigma^2$.

4 Amostragem simples

4.1 Atributos quantitativos

Na amostragem casual simples a amostra é obtida casualmente sobre toda a população, geralmente sem reposição. Este tipo de amostragem aplica-se a populações homogéneas ou, pelo menos, populações em que seja difícil "segregar" subpopulações mais homogéneas que a população considerada como um todo.

Sejam:

$X\ (\mu,\sigma^2)$ -- variável em estudo

$(X_1,X_2,...,X_n)$ --amostra aleatória de dimensão n

$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ -- média da amostra, estimador centrado de μ

$s_c^2 = s^2 \cdot \frac{n}{n-1} = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n-1}$ -- estimador centrado de σ^2

4.1.1 Intervalo de confiança

Na construção dos intervalos de confiança para a amostragem casual simples há que considerar separadamente, os casos em que se pode admitir que a população em estudo é normal daqueles em que esta hipótese não é aceitável.

4.1.1.1 Populações normais, grandes amostras

Se pudermos admitir que a população é normal, ou seja, que

$$X \cap \mathrm{N}\ (\mu,\sigma^2) \quad \rightarrow \quad Z = \frac{\overline{X}_n - \mu}{\sigma_{\overline{X}}} \cap \mathrm{N}\ (0,1)$$

é fácil chegar a um intervalo de confiança para μ :

$$P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(-z_{\alpha/2} \le \frac{\overline{X}_n - \mu}{\sigma_{\overline{X}}} \le z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(\overline{X}_n - \sigma_{\overline{X}}\ z_{\alpha/2} \le \mu \le \overline{X}_n + \sigma_{\overline{X}}\ z_{\alpha/2}\right) = 1-\alpha$$

Como já vimos atrás

$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$, na amostragem com reposição,

ou

$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}\ \frac{N-n}{N-1}$, na amostragem sem reposição.

Uma vez que σ^2 é geralmente desconhecido, podem utilizar-se os estimadores de $\sigma^2_{\overline{X}}$

$s^2_{\overline{X}} = \frac{s_c^2}{n}$, na amostragem com reposição,

ou

$s^2_{\overline{X}} = \frac{s_c^2}{n}\ \frac{N-n}{N}$, na amostragem sem reposição.

4.1.1.2 Populações normais, pequenas amostras

Se estivermos a trabalhar com uma pequena amostra (n<30), então $s^2_{\overline{X}}$ não é geralmente um bom estimador de $\sigma^2_{\overline{X}}$, pelo que se deve então recorrer à variável t com (n-1) graus de liberdade (g.l.):

$$\sqrt{n}\ \frac{\overline{X}_n - \mu}{s_c}\ \cap\ t_{n-1}$$

O intervalo de confiança fica então:

$$\mathrm{P}\left(\overline{X}_n - t_{\alpha/2}\frac{s_c}{\sqrt{n}} \le \mu \le \overline{X}_n + t_{\alpha/2}\frac{s_c}{\sqrt{n}}\right) = 1-\alpha$$

4.1.1.3 Populações não normais

Se não se puder admitir a hipótese de que X é $\mathrm{N}\ (\mu,\sigma^2)$, com base no teorema limite central podemos afirmar que $\overline{X}_n$ é assintóticamente $\mathrm{N}\ (\mu,\sigma^2/n)$. Pode assim construir-se o seguinte intervalo de confiança para $n \ge 30$:

$$\frac{\overline{X}_n - \mu}{\sigma_{\overline{X}}}\ \overset{a}{\cap}\ \mathrm{N}\ (0,1)$$

$$\downarrow$$

$$\mathrm{P}\ (\overline{X}_n - z_{\alpha/2}\ \sigma_{\overline{X}} \le \overline{X}_n + z_{\alpha/2}\ \sigma_{\overline{X}}) \approx 1-\alpha$$

Se σ^2 for desconhecido, utilizam-se os já referidos estimadores de $\sigma^2_{\overline{X}}$.

4.1.2 Cálculo da grandeza da amostra

A partir das expressões dos intervalos de confiança podemos deduzir as expressões para o cálculo da grandeza da amostra.

4.1.2.1 Grandes amostras

Se n > 30 (população normal ou não normal), temos que:

$$\mathrm{E} = z_{\alpha/2}\ \sigma_{\overline{X}} \quad \text{(erro absoluto)}$$

e

$$\mathrm{E}\% = \frac{z_{\alpha/2}\ \sigma_{\overline{X}}}{\overline{X}_n} \times 100 \quad \text{(erro percentual)}$$

Em aplicações práticas, as amostragens são geralmente planeadas em função de um erro percentual admissível. Uma vez fixado o erro percentual pode calcular-se o correspondente erro absoluto com base numa estimativa do valor médio da população:

$$\mathrm{E} = \frac{\mathrm{E}\%\ \overline{X}}{100}$$

Suponhamos que σ^2 é desconhecido e a amostragem feita sem reposição, como é usual. Então:

$$\mathrm{E} = z_{\alpha/2}\ \frac{s_c}{\sqrt{n}}\ \sqrt{1-\frac{n}{N}}$$

$$\mathrm{E}^2 = z^2_{\alpha/2}\ \frac{s_c^2}{n}\ \left(1-\frac{n}{N}\right)$$

$$\mathrm{E}^2 = z^2_{\alpha/2}\ \frac{s_c^2}{n} - z^2_{\alpha/2}\ \frac{s_c^2}{N}$$

Pode então resolver-se a equação em ordem a *n*:

$$n = \frac{s_c^2 \; z^2_{\alpha/2}}{\mathrm{E}^2 + \frac{z^2_{\alpha/2} \; s_c^2}{N}} = \frac{s_c^2 \; z^2_{\alpha/2}}{\left(\frac{\mathrm{E}\% \; \overline{X}}{100}\right) + \frac{z^2_{\alpha/2} \; s_c^2}{N}}$$

Para calcular *n* têm portanto que se estimar s_c^2 e $\overline{X}$, podendo recorrer-se a três processos:

1. Amostragens anteriores

2. Conhecendo a ordem de grandeza dos valores X_{max} e X_{min}

Com base nas estimativas do valor máximo e mínimo que podem ocorrer na população, é possível obter estimativas, embora um pouco "grosseiras" da média e do desvio padrão da população:

$$\overline{X} \approx \frac{X_{\max} + X_{\min}}{2}$$

Se admitirmos que $X \cap \mathrm{N} \; (\mu, \sigma^2)$ então

$$\mathrm{P} \; (X \in \underbrace{[\mu \pm 3\sigma]}_{6\sigma}) \approx 99\%$$

$$\downarrow$$

$$s \approx \frac{X_{\max} - X_{\min}}{4 \cdot 5}$$

Utiliza-se o denominador 4.5 em vez de 6, uma vez que existe evidência empírica de que as estimativas de σ obtidas por este processo são geralmente enviesadas por defeito.

3. Através de uma amostragem prévia que procuramos que seja representativa da população.

4.1.2.2 Pequenas amostras

No caso das pequenas amostras (população normal) temos que:

$$\mathrm{E} = t_{\alpha/2} \frac{s_c}{\sqrt{n}}$$

$$\mathrm{E}^2 = t^2_{\alpha/2} \frac{s_c^2}{n}$$

$$n = \frac{t^2_{\alpha/2} s_c^2}{\mathrm{E}^2} = \frac{t^2_{\alpha/2} s_c^2}{\left(\frac{\mathrm{E}\% \ \overline{X}}{100}\right)}$$

Como o valor de $t_{\alpha/2}$ depende de *n*, através do número de graus de liberdade *(n-1)*, não sabemos à partida o valor de *t* a introduzir na ezpressão de *n*. O cálculo deve ser feito por tentativas até que o valor obtido para *n* esteja em concordância com o número de graus de liberdade utilizado para o $t_{\alpha/2}$.

4.2 Atributos qualitativos

Analisemos agora a amostragem casual simples no caso de estrarmos interessados num atributo qualitativo. Este tipo de amostragem é particularmente importante para o inventário florestal, pelo facto de ser bastante utilizado para a estimativa das áreas dos diversos estratos presentes numa região para a qual não se dispomnha de cartografia actualizada. Esta estimativa é baseada na aplicação de uma grelha de pontos sobre toda a região, estimando-se as proporções de pontos que coincidem com cada um dos estratos de interesse. Uma vez calculados os intervalos de confiança para as proporções dos diversos estratos, estes são “convertidos” em intervalos de confiança para as áreas dos estratos por multiplicação de cada intervalo pela área total da região em estudo.

Sejam:

n – dimensão da amsotra

a – número de indivíduos da amostra com a característica C

$p^* = \frac{a}{n}$ proporção empírica, estimador centrado de p

$s^2 = p^* q^*$ variância empírica, estimador não centrado de pq

$s_c^2 = p^* q^* \frac{n}{n-1}$ estimador centrado de pq

4.2.1 Intervalo de confiança

Como já vimos $a \cap \mathrm{B}_i(n,p)$, com $\mu = np$ e $\sigma^2 = npq$

Para valores de np > 10 (podendo na prática admitir-se o valor 5) a distribuição binomial tende para a normal, sendo a convergência tanto mais rápida quanto mais próximo estiver p de 50%.

Temos então que

$$\frac{a-np}{\sqrt{npq}} \cap \mathrm{N}(0,1)$$

$$\downarrow$$

$$\frac{\frac{a}{n}-p}{\frac{\sqrt{npq}}{n}} = \frac{p^*-p}{\sqrt{pq}}\sqrt{n} \cap \mathrm{N}(0,1)$$

Resulta então o seguinte intervalo de confiança:

$$\mathrm{P}\left(p^* - z_{\alpha/2}\sqrt{\frac{pq}{n}} \le p \le p^* + z_{\alpha/2}\sqrt{\frac{pq}{n}}\right) \approx 1-\alpha$$

Neste intervalo aparece a variância exacta *pq*, pelo que, na prática, há que utilizar o seu estimador centrado $p^* q^* \frac{n}{n-1}$

Se *n*<100 e *p* estiver próximo de 0.10 ou 0.90, podemos recorrer ao seguinte resultado assimptótico (transformação angular de Fisher):

$$X = 2 \arcsin\sqrt{p^*} \underset{n\to\infty}{\cap} \mathrm{N}\left(2 \arcsin\sqrt{p}\ ,\ \frac{1}{n}\right)$$

4.2.2 Cálculo da grandeza da amostra

Neste caso o erro absoluto é dado por

$$\mathrm{E} = z_{\alpha/2}\sqrt{\frac{p^* q^*}{n-1}}$$

pelo que podemos deduzir facilmente que

$$n = \frac{z^2_{\alpha/2}\ p^* q^*}{\mathrm{E}^2} + 1 \approx \frac{z^2_{\alpha/2}\ p^* q^*}{\mathrm{E}^2}$$

Para "estimar" p^*q^* não deve utilizar-se uma amostragem prévia pois qualquer amostragem qualitativa exige um grande número de observações, especialmente se a proporção a avaliar for pequena. Podem surgir duas hipóteses:

1. Tem-se uma ideia da ordem de grandeza de p e não há qualquer problema em utilizar a expressão acima

2. Faz-se uma estimativa subjectiva de p. Ao contrário do que se passa com as grandezas quantitativas é por vezes possível estimar a ordem de grandeza da proporção de indivíduos com determinada característica

4.3 Exercícios

4.3.1 Amostragem casual simples quantitativa

Ex5.1 Amostragem da população I

a) Obtenha uma amostra de 48 indivíduos da população I, utilizando amostragem casual simples (caso já tenha resolvido o **Ex2.1** pode utilizar a amostra obtida na alínea a))

b) Utilize a amostra obtida em **a)** para estimar o volume ha^{-1} da população I a um nível de significância de 0.05. Calcule o erro percentual associado a esta estimativa.

c) Utilizando os resultados obtidos em **b)** por cada aluno, represente graficamente os desvios entre a média de cada amostra e a média da população ($\bar{\varepsilon} = (\bar{x} - \mu)$), os

desvios da média ($s_{\bar{x}}$) e os erros de amostragem ($t\ s_{\bar{x}}$) para cada amostra. No eixo dos *XX* deve representar as sucessivas amostras designadas por 1, 2, 3, ..., n (n = nº de alunos) e o eixo dos *YY* deve estar graduado em m^3ha^{-1}.

d) Calcule a grandeza da amostra necessária para estimar o volume ha^{-1} da população I com um erro percentual de 5%, a um nível de significância de 0.05, utilizando as seguintes estimativas de $\overline{X}$ e s^2:

- verdadeiro valor dos parâmetros (só a título de comparação; como é evidente, na prática estes valores não estão disponíveis)
- resultados de inventários anteriores (valores obtidos em **b)**)
- valores estimados por amostragem prévia (dimensão igual a 5)
- valores estimados com base na estimativa dos valores máximo e mínimo

Compare os resultados obtidos e procure interpretar as diferenças encontradas.

Ex5.2 Se resolver o problema **Ex5.1** em relação à população II, os resultados obtidos serão melhores ou piores? Justifique.

Ex5.3 Amostragem da população I por faixas

a) Considere a população I dividida em faixas verticais, cada uma constituída por 20 unidades casuais de 1000 m^2 (N=20, portanto). Obtenha uma amostra aleatória de 3 faixas (60 unidades de 1000 m^2) e utilize-a para estimar o volume ha^{-1} a um nível de significância de 0.05. Calcule o erro percentual associado a esta estimativa.

b) Como explica que tenha obtido, neste caso, um erro percentual superior ao obtido no problema 1, embora o esforço de amostragem (área total amostrada) seja superior?

4.3.2 Amostragem casual simples qualitativa

Ex5.4 Cálculo da grandeza da amostra para avaliação de áreas por amostragem qualitativa

a) Calcule o número de pontos de amostragem necessário para estimar a área da espécie em estudo na população IV, com um erro percentual (em termos de área) de 20% a um nível de significância de 0.05. Utilize as seguintes estimativas de p^*:

- verdadeiro valor de *p* (só a título de comparação; como é evidente, na prática estes valores não estão disponíveis)

- estimativa subjectiva de *p*

b) Obtenha uma amostra aleatória com dimensão igual à calculada em **a)** e utilize-a para estimar a área da espécie em estudo da população IV, a um nível de significância de 0.05. Calcule o erro percentual associado.

c) Utilizando os resultados obtidos por cada aluno faça gráficos equivalentes aos descritos no exercício **Ex5.1- c)** da amostragem casual simples.

Ex5.5 Utilizando a expressão para o cálculo da grandeza da amostra na amostragem qualitativa, represente graficamente a dimensão da amostra em função da proporção *p*, considerando erros percentuais de 5, 10, 15, e 20%.

Ex5.6 Cálculo da quadrícula para a estimação de áreas de estratos florestais delineados em fotografia aérea foto-interpretada.

a) Consultando os dados do Inventário Florestal realizado no concelho de Oliveira do Hospital, com uma área total de 23455 ha, no ano de 1992 (tabela 1), calcule o lado da quadrícula que deveria utilizar para a avaliação, com recurso a amostragem qualitativa sobre fotografia aérea de escala média 1:15000, a área ocupada por eucaliptais (povoamentos puros e mistos dominantes) com um erro percentual de 10%. Utilize um nível de significância de 0.05.

Tabela 6. Avaliação de áreas no Inventário de Oliveira do Hospital (Tomé *et al.*, 1992)

	Código nominal do estrato	Erro percentual (%)	Área ocupada (ha)	Proporção da área do concelho
Pinheiro bravo				
Povoamentos puros				
Regulares				
Nascedio e novedio	Pb0	20.9	288	1.23
Bastio	Pb1	26.9	174	0.74
Fustadio e alto fuste	Pb2	8.5	1640	6.99
Total de regulares		7.4	2102	8.96
Irregulares				
Cobertura > 70%	Pb>70%	8.1	1806	7.70
Cobertura 40 a 70%	Pb 40-70%	10.3	1149	4.90

Cobertura < 40%	Pb<40%	17.5	410	1.75
Total de irregulares		5.7	3365	14.35
Total de povoamentos puros		4.2	5467	23.31
Povoamentos mistos dominantes				
P. bravo com P. manso	Pb+Pm	11.0	1023	4.32
P. bravo com Folhosas	Pb+Dx	18.8	357	1.52
Total de mistos dominantes		9.4	1370	5.84
Total de Pinheiro bravo		**3.6**	**6837**	**29.15**
Pinheiro manso				
Povoamentos puros	Pm	31.3	129	0.55
Total de povoamentos puros				
Povoamentos mistos dominantes				
P. manso com P. bravo	Pm+Pb	13.2	713	3.04
P. manso com Folhosas	Pm+Dx	42.7	70	0.30
Total de mistos dominantes		12.5	783	3.34
Total de Pinheiro manso		**11.6**	**912**	**3.89**
Eucalipto globulus				
Povoamentos puros	Eg	23.4	230	0.98
Total de povoamentos puros				
Povoamentos mistos dominantes				
Eucalipto com P. bravo	Eg+Pb	67.2	28	0.12
Total de mistos dominantes		22.1	258	1.10
Total de Eucalipto globulus				
Folhosas diversas				
Total de Folhosas diversas	Dx	21.8	271	1.16
Árvores dispersas				
Total de Árvores dispersas	k	19.9	316	1.35
Total de Área Florestal		3.1	8594	36.64

b) No inventário florestal da zona Norte Litoral realizado no âmbito da 2ª revisão do Inventário florestal Nacional, as áreas dos diversos estratos florestais foram estimadas com base em fotografia aérea captada em 1979. Para os estratos de pinheiro bravo e eucalipto (povoamentos puros e mistos dominantes) foram apuradas as áreas de 157000 e 7300 ha, respectivamente. Sabendo que a área dos distritos incluídos nesta zona (Viana do Castelo, Braga e Porto) é de 725000 ha, calcule o lado da quadrícula adequada para amostrar novamente essa zona com o objectivo de actualizar as áreas dos estratos referidos. Pretendem-se erros de amostragem inferiores a 10% em ambos os estratos. Utilize um nível de significância de 0.05.

5 Amostragem estratificada

5.1 Atributos quantitativos

Se uma população é composta de outras subpopulações mais homogéneas, o cálculo dos parâmetros faz-se de maneira a separar os vários componentes e, consequentemente, a variância total reduz-se numa quantidade correspondente à variação entre componentes. Este processo de dividir a população original em subpopulações mais homogéneas com o objectivo de reduzir a variância chama-se "estratificação" e cada subpopulação segregada é um "estrato".

A estratificação pode ser feita essencialmente de duas maneiras:

- se a causa da heterogeneidade é conhecida (p.e. exposições, situações na encosta, declives, linhas de água, etc.), a estratificação é feita de acordo com estes factores e aqui desempenha um papel preponderante a fotografia aérea

- em zonas onde é difícil fazer o reconhecimento directo da área a amostrar (p.e. florestas tropicais) e não se disponha de fotografia aérea, apenas é possível dividir a zona em figuras geométricas regulares, designadas então por blocos. Consegue-se geralmente assim alguma redução da variância, visto que os indivíduos vizinhos apresentam mais semelhanças entre si do que com outros distantes.

Como já vimos, a população I é nitidamente estratificada. Se tivermos recurso a cartografia ou a fotografia aérea podemos "optimizar a estratificação da população, dividindo-a em três estratos de acordo com a densidade do povoamento ou qualquer outra variável que ocasione a variação espacial que é visível nos volumes. A figura 13 representa uma estratificação "optimizada" da população I. Outro modo de estratificar a população I, será dividi-la em 16 blocos, tal como se pode ver na mesma figura, designados pelo conjunto (letra, letra romana). Por exemplo, o bloco superior esquerdo será designado por (A I). Veremos em seguida que, mesmo a divisão da população em blocos é capaz de induzir uma redução bastante significativa na variância da população, embora, como é óbvia, esta redução seja mais evidente com a optimização da estratificação.

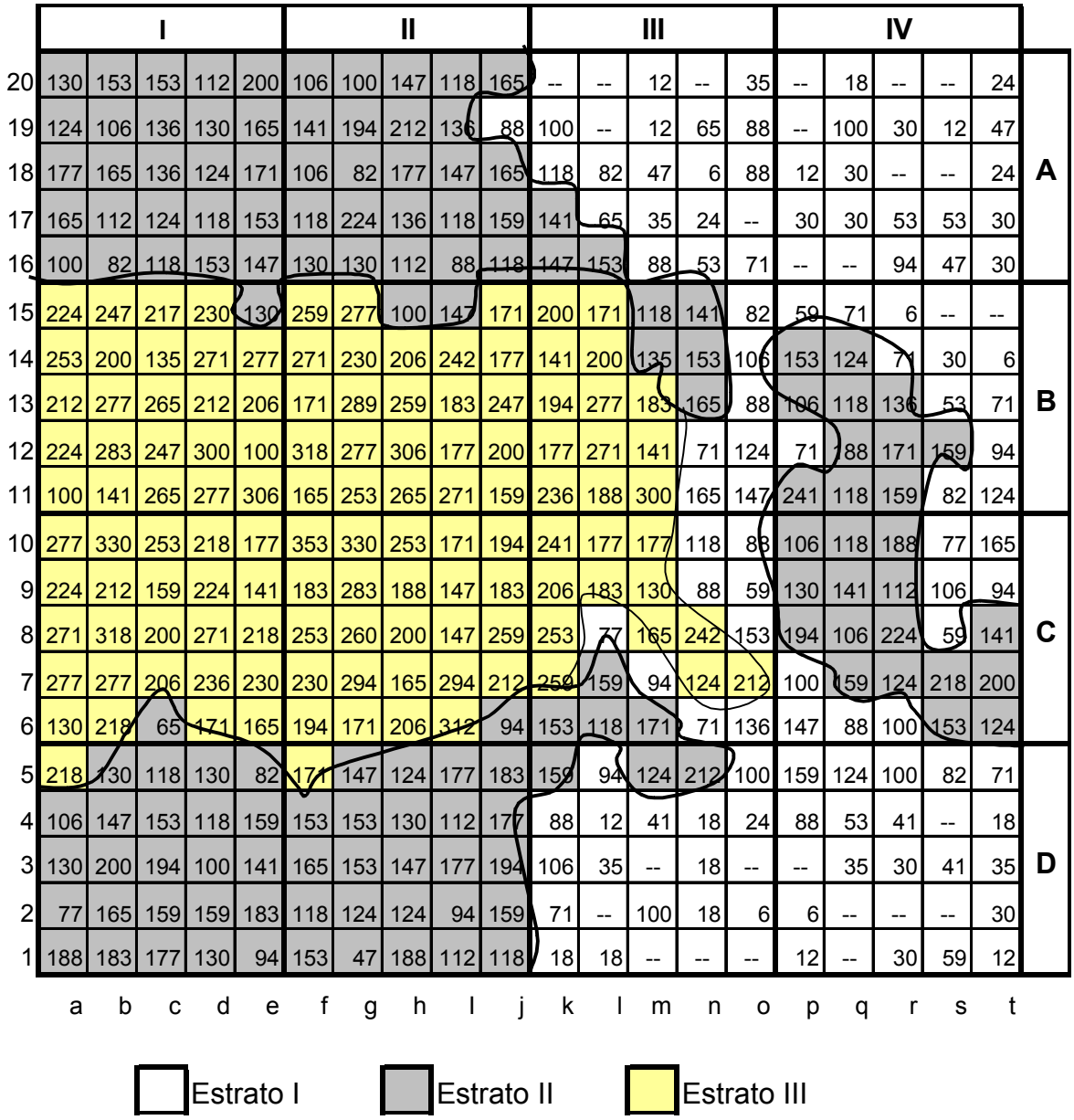


Figura 13. Talhão de uma Mata Nacional (população I do exemplo 1), após estratificação (população I.b)

5.1.1 Efeitos da estratificação na média e na variância da população

Sejam:

M -- número de estratos

N_j -- número de indivíduos por estrato ($j=1,...,M$)

x_{ij} -- observação x na parcela i do estrato j ($i=1,...,N_j$)

$$\mu_j = \frac{\sum_{i=1}^{N_j} x_{ij}}{N_j}$$ -- média do estrato *j*

$$\sigma_j^2 = \frac{\sum_{i=1}^{N_j} (x_{ij} - \mu_j)^2}{N_j}$$ -- variância do estrato *j*

$$N = N_1 + N_2 + ... + N_M = \sum_{j=1}^{M} N_j$$ -- número total de indivíduos na população

$$P_j = \frac{N_j}{N} = \frac{N_j}{\sum_{j=1}^{M} N_j}$$ -- proporção do estrato *j* na população $\left(\sum_{j=1}^{M} P_j = 1 \right)$

Podemos então calcular a média e a variância da população estratificada, ponderando as médias e as variâncias de cada estrato com a proporção respectiva.

Média:

$$\mu_{st} = \sum_{j=1}^{M} P_j \mu_j = \sum_{j=1}^{M} \frac{N_j}{N} \mu_j = \sum_{j=1}^{M} \frac{N_j}{N} \frac{1}{N_j} \sum_{i=1}^{N_j} x_{ij} = \frac{\sum_{j=1}^{M} \sum_{i=1}^{N_j} x_{ij}}{N} = \frac{\sum_{k=1}^{N} x_k}{N} = \mu$$

Variância:

$$\sigma_{st}^2 = \sum_{j=1}^{M} P_j \sigma_j^2 = \sum_{j=1}^{M} \frac{N_j}{N} \sigma_j^2 = \sum_{j=1}^{M} \frac{N_j}{N} \frac{1}{N_j} \sum_{i=1}^{N_j} (x_{ij} - \mu_j)^2 = \frac{\sum_{j=1}^{M} \sum_{i=1}^{N_j} (x_{ij} - \mu_j)^2}{N} \neq \sigma^2$$

Como fácilmente se conclui das expressões acima, a média é idêntica à da população não estratificada. Com a variância o caso é completamente diferente. a variância já não se compõe dos quadrados dos desvios das observações para amédia comum μ, mas sim para a média μ_j do estrato respectivo. É fácil verificar que:

$$\sum_{j=1}^{M}\sum_{i=1}^{N_j}(x_{ij}-\mu)^2 \geq \sum_{j=1}^{M}\sum_{i=1}^{N_j}(x_{ij}-\mu_j)^2$$

verificando-se a igualdade só se $\mu_j = \mu \qquad \forall j$

Vejamos em que medida é que a variância é reduzida pela estratificação:

$$x_{ij}-\mu = x_{ij}-\mu+\mu_j-\mu_j$$

$$(x_{ij}-\mu)^2 = \left[(x_{ij}-\mu_i)+(\mu_j-\mu)\right]^2$$

$$(x_{ij}-\mu)^2 = (x_{ij}-\mu_i)^2 + 2\ (x_{ij}-\mu_i)(\mu_j-\mu)+(\mu_j-\mu)^2$$

$$\sum_{j=1}^{M}\sum_{i=1}^{N_j}(x_{ij}-\mu)^2 = \sum_{j=1}^{M}\sum_{i=1}^{N_j}(x_{ij}-\mu_j)^2 + 2\ \sum_{j=1}^{M}(\mu_j-\mu)\underbrace{\sum_{i=1}^{N_j}(x_{ij}-\mu_j)}_{0}+\sum_{j=1}^{M}\sum_{i=1}^{N_j}(\mu_j-\mu)^2$$

$$\underbrace{\sum_{j=1}^{M}\sum_{i=1}^{N_j}(x_{ij}-\mu)^2}_{\text{variação total}} = \underbrace{\sum_{j=1}^{M}\sum_{i=1}^{N_j}(x_{ij}-\mu_j)^2}_{\text{variação da população estratificada}} + \underbrace{\sum_{j=1}^{M}N_j(\mu_j-\mu)^2}_{\text{variação entre estratos}}$$

Dividindo ambos os membros por N, obtém-se expressão idêntica mas em termos das variâncias:

$$\sigma^2 = \sigma_{st}^2 + \sum_{j=1}^{M}P_j(\mu-\mu_j)^2$$

$$\sigma_{st}^2 = \sigma^2 - \sum_{j=1}^{M}P_j(\mu-\mu_j)^2 = \sum_{j=1}^{M}P_j\sigma_j^2$$

Na prática, σ_{st}^2 pode ser calculado com base na seguinte expressão:

$$\sigma_{st}^2 = \frac{1}{N}\left(\sum_{i=1}^{M}\sum_{i=1}^{N_j} x_{ij}^2 - \sum_{j=1}^{M}\mu_j \sum_{i=1}^{N_j} x_{ij}\right)$$

Se considerarmos o estrato como um factor com *M* níveis, estes resultados podem exprimir-se em termos de uma tabela de análise de variância (tabela **).

Tabela 7. Análise de variância correspondente a uma amostragem estratificada

Origem da variação	g.l.	Somas de quadrados	Quadrados médios
Dentro dos estratos	*N*-*M*	$\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(x_{ij}-\mu_j\right)^2$	S_{st}^2
Entre estratos	*M*-1	$\sum_{j=1}^{M} N_j\left(\mu_j-\mu\right)^2$	S_{est}^2
Total	*N*-1	$\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(x_{ij}-\mu\right)^2$	$S^2 = \frac{(N-M)S_{st}^2 + (M-1)S_{est}^2}{N-1}$

No caso de uma amostra, a partição da soma de quadrados total nas suas duas componentes só se mantém se a fracção de amostragem for igual em todos os estratos e se as variâncias dentro de cada estrato não diferirem significativamente. Neste caso, e com base na teoria da análise de variância, pode utilizar-se o quadrado médio dentro dos estratos como estimador da variância comum dentro dos estratos:

$$\hat{\sigma}_{st}^2 = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left(x_{ij}-\overline{X_j}\right)^2}{n-M}$$

Para esclarecer melhor o efeito da estratificação na média e na variância, consideremos a populaçãp 1 dividida nos 16 (=*M*) blocos A I a D IV. Cada um dos blocos tem 25 (=N_j) parcelas. Queremos agora calcular a variância da população estratificada e fazer a correspondente análise de variância, comparando os resultados (variância, desvio padrão e coeficiente de variação) com os correspondentes valores da população não estratificada.

Tabela 8. Elementos para a análise de variância da população I estratificada em 16 blocos de igual dimensão

Bloco					
Número	Código	μ_j	$\sum_{i=1}^{N_j}(x_{ij}-\mu_j)^2$	$N_j\ (\mu_j-\mu)^2$	$\sum_{i=1}^{N_j}(x_{ij}-\mu)^2$
1	A I	138.2	18101.36	74.4	18175.75
2	A II	136.6	33716.16	0.4	33716.55
3	A III	57.2	58842.00	156954.6	215796.63
4	A IV	26.6	18400.16	301812.9	320213.05
5	B I	224.0	87548.96	191515.6	279064.60
6	B II	224.8	77294.00	195209.3	272503.33
7	B III	167.0	83394.96	23294.4	106689.35
8	B IV	96.4	95438.16	39990.0	135428.16
9	C I	218.7	88123.04	169270.5	257393.57
10	C II	223.0	98476.96	187510.7	285987.61
11	C III	154.2	83089.36	7854.4	90943.75
12	C IV	135.0	46428.96	54.4	46483.35
13	D I	145.6	34675.76	2118.3	36794.06
14	D II	144.0	27666.00	1430.7	1163.87
15	D III	50.4	80900.00	185050.5	265950.53
16	D IV	41.4	43388.00	225791.3	269179.28
	$\sum_{j=1}^{M}$		975483.84	1687932.47	2635483.45

$$\sigma^2_{str} = \frac{975483.84}{400} = 2438.71$$

$$\sigma_{str} = \pm\ 49.38\ \ m^3\ ha^{-1}$$

$$C.V. = \sigma_{str}\,\% = \frac{\pm\ 49.38}{136.44} \times 100 = \pm\ .36.19\%$$

Com os dados da tabela **, facilmente se obtém a seguinte partição da variância:

Variação dentro dos blocos	975 483.84
Variação entre blocos	1 687 932.47
Variação total	2 635 483.45

Neste exemplo, embora a estratificação não tenha sido optimizada, a maior parte da variação (1 687 932.47, ou seja, cerca de 64%) é causada pela variação entre blocos. Na população não estratificada o CV foi de 59.81%, enquanto que na população estratificada se reduziu a 36.19%, pelo que ceca de 40% do desvio padrão total da população pode ser atribuído às diferenças entre blocos.

5.1.2 Intervalo de confiança

A amostragem estratificada incide, como o próprio nome indica, numa população estratificada. Obtém-se uma amostra casual e independente de cada estrato. A independência das amostras obtidas nos diferentes estratos é bastante importante, uma vez que na dedução das expressões da variância se assume que é possível obter a variância total como a soma das variâncias dos vários estratos, o que só é possível se houver independência das amostras obtidas nos vários estratos.

Sejam:

$(X_{1j},X_{2j},...,X_{n_j j})$ -- amostra aleatória de dimensão n_j para o estrato j

$\overline{X}_j = \overline{X}_{nj} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}$ -- média da amostra do estrato *j*, estimador centrado de μ_j

$s^2_{j,c} = s^2_j \frac{n_j}{n_j - 1} = \frac{1}{n_j - 1} \sum_{i=1}^{n_j}(x_{ij} - \overline{X}_j)^2$ -- estimador centrado de σ^2_j

Como já vimos, o parâmetro média a estimar é neste caso

$$\mu = \sum_{j=1}^{M} P_j \mu_j$$

A média da amostra

$$\overline{X} = \frac{\sum_{j=1}^{M}\sum_{i=1}^{n_j} x_{ij}}{n} = \sum_{j=1}^{M} p_j \overline{X}_j = \sum_{j=1}^{M} \frac{n_j}{n} \overline{X}_j$$

não é, em geral, um estimador centrado de μ, uma vez que a intensidade de amostragemem cada estrato (p_i) nem sempre é igual à proporção que o estrato representa na população total.

Utiliza-se então o seguinte estimador centrado da média da população:

$$\overline{X}_{st} = \sum_{j=1}^{M} P_j \overline{X}_j$$

5.1.2.1 Populações normais, grandes amostras

Se cada um dos M estratos for aproximadamente normal, então

$$\overline{X}_{st} = \sum_{j=1}^{M} P_j \overline{X}_j$$

sendo uma combinação linear de normais, tem uma distribuição normal de parâmetros μ e $\sigma^2_{\overline{X}_{st}}$

Vejamos a que é igual $\sigma^2_{\overline{X}_{st}}$:

$$\sigma^2_{\overline{X}_{st}} = \text{var}\left(\sum_{j=1}^{M} P_j \overline{X}_j \right)$$

$$= \sum_{j=1}^{M} P_j^2 \sigma^2_{\overline{X}_j} = \sum_{j=1}^{M} P_j^2 \frac{\sigma_j^2}{n_j}$$

Assim:

$$\frac{\overline{X}_{st} - \mu}{\sigma_{\overline{X}_{st}}} \cap \text{N}(0,1)$$

Podemos então construir o intervalo de confiança:

$$P\left(\overline{X}_{st} - z_{\alpha/2} \sigma_{\overline{X}_{st}} \leq \mu \leq \overline{X}_{st} + z_{\alpha/2} \sigma_{\overline{X}_{st}} \right) = 1 - \alpha$$

Uma vez que as variâncias dentro de cada estrato (σ_j^2) são geralmente desconhecidas, há então que encontrar um estimador para $\sigma_{\overline{X}_{st}}$

Já vimos que $\sigma^2_{\overline{X}_j}$ pode ser estimado por

$$s^2_{\overline{X}_j} = \frac{s^2_{j,c}}{n_j}\left(1 - f_j\right)$$

Temos assim para estimador de $\sigma^2_{\overline{X}_{st}}$

$$s^2_{\overline{X}_{st}} = \sum_{j=1}^{M} P_j^2 \frac{s^2_{j,c}}{n_j}\left(1 - f_j\right)$$

$$= \frac{1}{N^2}\sum_{j=1}^{M} N_j(N_j - n_j)\frac{s^2_{j,c}}{n_j}$$

$$= \frac{1}{N^2}\sum_{j=1}^{M} \frac{N_j(N_j - n_j)}{n_j(n_j - 1)}\sum_{i=1}^{n_j}\left(x_{ij} - \overline{X}_j\right)^2$$

Embora na prática se costume utilizar este estimador com poucas unidades por estrato (por vezes apenas duas) é desejável "pesar" bastante mais a amostragem pois só com n_j próximo de 30 podemos, com segurança, substituir σ_j^2 por $S^2_{j,c}$.

5.1.2.2 Populações normais, pequenas amostras

No caso de pretendermos utilizar uma pequena amostra há que admitir que σ_j^2 tem o mesmo valor (ou semelhante) em todos os estratos. Com base na análise de variância da amostra utiliza-se, então, o quadrado médio dentro dos estratos (ou quadrado médio de erro) como estimativa da variância comum σ_{st}^2:

$$s^2_{st,c} = \underbrace{\frac{1}{n}\sum_{j=1}^{M}\sum_{i=1}^{n_j}\left(x_{ij} - \overline{X_j}\right)^2}_{s^2_{st}} \frac{n}{n - M}$$

Nestas situações a amostragem é, geralmente, proporcional pelo que $\sigma^2_{\overline{X}_{st}}$ e o seu estimador $s^2_{\overline{X}_{st}}$ tomam as seguintes formas simplificadas:

$$\sigma^2_{\overline{X}_{st}} = \frac{\sigma^2}{n}$$

$$s^2_{\overline{X}_{st}} = \frac{s^2_{st,c}}{n}(1-f)$$

Chama-se a atenção para o facto de este procedimento só ser válido se as variâncias dos estratos não diferirem significativamente de estrato para estrato, o que pode ser confirmado, por exemplo, com o teste de Bartlett.

Como é sabido da teoria da análise da variância:

$$\frac{ns^2_{st}}{\sigma^2} = \sum_{j=1}^{M}\sum_{i=1}^{n_j}\left(\frac{x_{ij}-\overline{X}_j}{\sigma}\right)^2 \cap \chi^2_{n-M}$$

Por outro lado, já vimos que se a população em cada um dos estratos for aproximadamente normal, então

$$\frac{\overline{X}_{st}-\mu}{\sigma_{\overline{X}_{st}}} = \sqrt{n}\frac{\overline{X}_{st}-\mu}{\sigma} \cap \mathrm{N}(0,1)$$

Resulta então que

$$t = \frac{\sqrt{n}\dfrac{\overline{X}_{st}-\mu}{\sigma}}{\sqrt{n}\dfrac{s_{st}}{\sigma}\ \dfrac{1}{\sqrt{n-M}}} = \sqrt{n-M}\ \frac{\overline{X}_{st}-\mu}{s_{st}} \cap t_{n-M}$$

O intervalo de confiança fica portanto:

$$P\left(\overline{X}_{st} - t_{\alpha/2}\frac{s_{st,c}}{\sqrt{n}} \le \mu \le \overline{X}_{st} + t_{\alpha/2}\frac{s_{st,c}}{\sqrt{n}}\right) = 1-\alpha$$

Note-se que o recurso à variável t de Student implica que admitimos que a população em estudo é normal, embora na prática este intervalo de confiança seja utilizado generalizadamente sempre que estamos a trabalhar com pequenas amostras. Há contudo que ter em atenção que, ao aplicarmos este intervalo de confiança em populações não normais, não temos a garantia de que as estimativas do erro de amostragem estejam correctas.

5.1.2.3 Populações não normais

Se n > 30, o teorema limite central permite afirmar que, mesmo que não seja razoável admitir a normalidade da população, $\overline{X}_{st}$ é aproximadamente normal de parâmetros μ e $\sigma^2_{\overline{X}st}$ pelo que se pode construir o seguinte intervalo de confiança:

$$P\left(\overline{X}_{st} - z_{\alpha/2}\,\sigma_{\overline{X}_{st}} \le \mu \le \overline{X}_{st} + z_{\alpha/2}\,\sigma_{\overline{X}_{st}}\right) \approx 1-\alpha$$

Também neste caso há que estimar $\sigma^2_{\overline{X}st}$ tal como se indicou no caso das populações normais, grandes amostras.

5.1.3 Cálculo da grandeza da amostra

No cálculo da grandeza da amostra na amostragem estratificada, há que fixar não só n (grandeza total da amostra) como também os vários n_j (grandeza da amostra no estrato *j*). Podem seguir-se vários critérios:

- Amostragem proporcional à dimensão dos estratos
- Minimização do erro para um determinado custo
- Minimização do custo para um determinado erro
- Amostragem óptima (ou de Neyman)

Relembremos que se fixarmos um erro percentual $\mathrm{E}\%$, então

$$E = \frac{E\%}{100}\ \overline{X}_{st}$$

A estimação de valores $s^2_{j,c}$ e $\overline{X}_{st}$, necessários nas várias fórmulas para o cálculo da grandez da amostra, é feita, em qualquer destes critérios, pelos processos já apontados na amostragem casual simples.

5.1.3.1 Grandes amostras

Amostragem proporcional à dimensão dos estratos

Para este tipo de amostragem obtemos a fórmula

$$P_j = \frac{A_j}{A} = \frac{N_j}{\mathrm{N}} = \frac{n_j}{n} \qquad (A \text{ – área total; } A_j \text{ – área do estrato } j)$$

$$\downarrow$$

$$n_j = P_j \ .n$$

Pode deduzir-se uma expressão para o cálculo da grandeza da amostra:

$$s^2_{\overline{X}st} = \sum_{j=1}^{M} P \frac{\mathrm{P}_j^2}{n_j} s^2_{j,c}\left(1 - f_j\right)$$

$$\frac{n_j}{n} = \frac{N_j}{N} \quad \rightarrow \quad n_j = P_j \cdot n \quad \rightarrow \quad f_j = \frac{n_j}{N_j} = \frac{P_j n}{P_j N} = \frac{n}{N}$$

$$s^2_{\overline{X}st} = \frac{1}{n}\left(1 - \frac{n}{\mathrm{N}}\right) \sum_{j=1}^{M} P_j s_{j,c}$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{j=1}^{M} \mathrm{P}_j s^2_{j,c}$$

$$= \frac{1}{n}\sum_{j=1}^{M} P_j s^2_{j,c} - \frac{1}{N}\sum_{j=1}^{M} P_j s^2_{j,c}$$

Podemos agora reescrever a expressão doerro absoluto e elevá-la ao quadrado:

$$\mathrm{E}^2 = s^2_{\overline{X}st} z^2_{\alpha/2}$$

$$= \frac{z^2_{\alpha/2}}{n}\sum_{j=1}^{M} P_j s^2_{j,c} - \frac{z^2_{\alpha/2}}{N}\sum_{j=1}^{M} P_j s^2_{j,c}$$

Finalmente, obtém-se a expressão para o cálculos da grandeza da amostra:

$$n = \frac{z^2_{\alpha/2} \sum_{j=1}^{M} P_j s^2_{j,c}}{\mathrm{E}^2 + \frac{z^2_{\alpha/2} \sum_{j=1}^{M} P_j s^2_{j,c}}{N}}$$

Minimização do erro para um determinado custo

A função de custo mais simples é do tipo

$$C = C_o + \sum_{j=1}^{M} C_j n_j$$

$C_o -$ custos gerais

$C_j -$ custos de uma unidade no estrato j

Os custos gerais são aqueles que estão implícitos à realização da amostragem, independentemente do número de indivíduos que se venham a seleccionar para fazer parte da amostra. Para além destes custos são geralmente muito importantes os custos associados à medição/avaliação de cada indivíduo que é seleccionado para fazer parte da amostra.

Esta função é apropriada quando o maior contributo para o custo é dado pela medição das unidades de campo. Vários estudos empíricos têm no entanto sugerido que, se o custo de acesso às unidades for substancial, este custo se representa melhor pela expressão

$\sum_{j=1}^{M} t_j \sqrt{n_j}$ onde t_j é o custo de acesso por unidade.

Consideraremos apenas a função linear do custo.

Reescrevamos a expressão do erro absoluto de amostragem numa forma em que se isole o valor n_j, de modo a ser mais fácil deduzir a espressão para o cálculo da grandeza da amostra:

$$\mathrm{E} = \sigma_{\overline{X}_{st}} z_{\alpha/2}$$

com $\sigma^2_{\bar{X}_{st}} = \sum_{j=1}^{M} P_j^2 \frac{s_{j,c}^2}{n_j}\left(1-f_j\right)$

$$= \sum_{j=1}^{M} P_j^2 \frac{s_{j,c}^2}{n_j} - \sum_{j=1}^{M} P_j^2 \frac{s_{j,c}^2}{n_j}\frac{n_j}{N_j}$$

$$= \sum_{j=1}^{M} P_j^2 \frac{s_{j,c}^2}{n_j} - \sum_{j=1}^{M} P_j^2 \frac{s_{j,c}^2}{N_j}$$

Vem então:

$$\mathrm{E}^2 = z_{\alpha/2}^2 \sum_{j=1}^{M} P_j^2 \frac{s_{j,c}^2}{n_j} - z_{\alpha/2}^2 \sum_{j=1}^{M} P_j^2 \frac{s_{j,c}^2}{N_j}$$

Há que minimizar esta expressão sujeita à condição

$$\sum_{j=1}^{M} C_j n_j + C_0 - C = 0$$

Utilizando o método dos multiplicadores de Lagrange, vem que:

$$\frac{\delta}{\delta n_j}\left(\mathrm{E}^2 + \lambda\left(\sum_{j=1}^{M} C_j n_j + C_0 - C\right)\right) = 0$$

Para um determinado estrato n_j teremos

$$-z_{\alpha/2}^2 P_j^2 \frac{s_{j,c}^2}{n_j^2} + \lambda C_j = 0$$

$$-z_{\alpha/2}^2 P_j^2 s_{j,c}^2 + \lambda C_j n_j^2 = 0$$

$$n_j\sqrt{\lambda} = \frac{z_{\alpha/2} P_j s_{j,c}}{\sqrt{C_j}}$$

Para todos os estratos obtemos:

$$\sum_{j=1}^{M} n_j \sqrt{\lambda} = n\sqrt{\lambda} = \sum_{j=1}^{M} \frac{z_{\alpha/2} P_j s_{j,c}}{\sqrt{C_j}}$$

Relacionando então n_j com n:

$$\frac{n_j}{n} = \frac{\dfrac{P_j s_{j,c}}{\sqrt{C_j}}}{\displaystyle\sum_{j=1}^{M}\left(\frac{P_j s_{j,c}}{\sqrt{C_j}}\right)} \qquad \Rightarrow \qquad n_j = \frac{\dfrac{P_j s_{j,c}}{\sqrt{C_j}}}{\displaystyle\sum_{j=1}^{M}\left(\frac{P_j s_{j,c}}{\sqrt{C_j}}\right)}\ n$$

Esta expressão leva às seguintes regras gerais de conduta. Num determinado estrato, tome-se uma amostra maior se:

- o estrato é maior
- o estrato é internamente mais variável
- a amostragem é mais barata neste estrato

Basta agora substituir os valores óptimos de n_j na expressão do custo e resolver em ordem a n, para obter a expressão que dá o número total de unidades a amostrar:

$$n = \frac{(C - C_0)\displaystyle\sum_{j=1}^{M}\left(\frac{N_j s_{j,c}}{\sqrt{C_j}}\right)}{\displaystyle\sum_{j=1}^{M}\left(N_j s_{j,c}\sqrt{C_j}\right)}$$

<u>Minimização do custo para um determinado erro</u>

Neste caso há que minimizar a expressão do custo sujeita à condição

$$\mathrm{E}^2 - z_{\alpha/2}^2 \sum_{j=1}^{M} P_j^2 \frac{s_{j,c}^2}{n_j} + z_{\alpha/2}^2 \sum_{j=1}^{M} P_j^2 \frac{s_{j,c}^2}{N_j} = 0$$

Utilizando, mais uma vez, o método dos multiplicadores de Lagrande é fácil provar que a expressão para o cálculo de n_j é a mesma que no caso da minimização do erro para um determinado custo.

Para obter n, substitui-se a expressão de n_j na expressão do erro e explicita-se em ordem a n, obtendo:

$$n = \frac{\left(\sum_{j=1}^{M} P_j s_{j,c} \sqrt{C_j}\right)\left(\sum_{j=1}^{M} \frac{P_j s_{j,c}}{\sqrt{C_j}}\right) z_{\alpha/2}^2}{\mathrm{E}^2 + \frac{1}{N}\sum_{j=1}^{M} P_j s_{j,c}^2 z_{\alpha/2}^2}$$

Amostragem óptima (ou de NEYMAN)

Um caso importante é aquele em que $C_j = C$, isto é, o custo para a medição/avaliação de um indivíduo é o mesmo para todos os estratos. O custo torna-se $C = C_0 + Cn$ e a amostragem óptima para custo fixo reduz-se à amostragem óptima (erro mínimo) para determinada dimensão da amostra (n).

Temos então

$$n_j = \frac{P_j s_{j,c}}{\sum_{j=1}^{M} P_j s_{j,c}} \quad \Leftrightarrow \quad n = \frac{N_j s_{j,c}}{\sum_{j=1}^{M} N_j s_{j,c}} n$$

A fórmula que fornece a grandeza total da amostra resulta, uma vez mais, da substituição do valor de n_j na expressão do erro

$$n = \frac{\left(\sum_{j=1}^{M} P_j s_{j,c}\right)^2 z_{\alpha/2}^2}{\mathrm{E}^2 + \frac{z_{\alpha/2}^2 \sum_{j=1}^{M} P_j s_{j,c}^2}{N}}$$

5.1.3.2 Pequenas amostras

Uma vez que a amostragem estratificada com pequenas amostras se baseia na hipótese de que as variâncias dentro de cada estrato não diferem significativamente, utiliza-se, neste caso, sempre uma amostragem proporcional à dimensão dos estratos. Há então que utilizar a seguinte expressão:

$$n = \frac{t^2_{\alpha/2}\, s^2_{st,c}}{\mathrm{E}^2}$$

Tal como no caso da amsotragem casual simples, o valor de $t_{\alpha/2}$ depende de *n*, através do número de graus de liberdade *(n-1)*, pelo que não sabemos à partida o valor de *t* a introduzir na ezpressão de *n*. O cálculo deve ser feito por tentativas até que o valor obtido para *n* esteja em concordância com o número de graus de liberdade utilizado para o $t_{\alpha/2}$.

5.2 Exercícios

Ex6.1 Prove a igualdade

$$\sum_{j=1}^{M}\sum_{i=1}^{N_j}(X_{ij} - \mu_j)^2 = \sum_{j=1}^{N_j} X_{ij}^2 - \sum_{j=1}^{M}\mu_j \sum_{i=1}^{N_j} X_{ij}$$

Ex6.2 Análise de variância de populações estratificadas

a) Faça a análise de variância e cálculo de σ_{str} e σ_{str}% para a população I dividida em 16 blocos de igual dimensão.

b) Idêntico a **a)**, mas em relação à população II.

c) Compare os resultados obtidos em **a)** e **b)** e procure explicá-los

Ex6.3 O mesmo que em **Ex6.2**, mas em relação à população I.b dividida nos seus 3 estratos (ficheiro POP-I.b.XLS).

Compare com **Ex6.2 - a)** e **b)** e interprete os resultados

Ex6.4 Estimadores na amostragem estratificada

a) Verifique que

$$\overline{X} = \sum_{j=1}^{M} \frac{n_j}{nj} \overline{X_j}$$

não é, em geral, um estimador centrado de μ_i..

b) Em que caso particular se verifica esta condição?

c) Verifique também que

$$\overline{X_{st}} = \sum_{j=1}^{M} P_j \overline{X_j}$$

é um estimador centrado de μ.

Ex6.5 Encontre expressões simples para $s^2_{\overline{x}_{st}}$, nos seguintes casos particulares:

a) $f_j < 0.1$ em todos os estratos

b) Amostragem proporcional à dimensão dos estratos

c) Amostragem proporcional e variância dentro dos estratos não diferindo significativamente.

Ex6.6 Qual a razão porque se utiliza, em geral, uma amostragem proporcional à dimensão dos estratos quando a variância tem o mesmo valor (ou semelhante) em todos os estratos?

Ex6.7 Verifique que, no caso de populações não normais, $\overline{X_{st}}$ é aproximadamente normal para $n \geq 30$.

Ex6.8 Se fosse possível dividir uma população em estratos de tal modo que todos os indivíduos do mesmo estrato fossem iguais, qual seria o erro associado à estimativa de μ?

Ex6.9 É vulgar utilizar-se uma amostragem por blocos = estratos de igual dimensão (se a população tiver uma forma regular). Neste caso, muitas das formas se simplificam. Deduza as fórmulas para o cálculo de $s^2_{\bar{x}_{st}}$ e para o cálculo da grandeza da amostra:

a) No caso geral

b) Se a amostragem for proporcional

c) Se a amostragem for proporcional e as variâncias dentro dos estratos não diferirem significativamente.

Ex6.10 Amostragem da população I estratificada por blocos (proporcional à dimensão dos estratos)

a) Obtenha uma amostra da população I, de dimensão igual a 48, utilizando uma amostragem estratificada em blocos de igual dimensão (16 blocos). Distribua a amostra proporcionalmente à dimensão dos estratos.

b) Utilize a amostra obtida em **a)** para estimar o volume ha^{-1} da população I, a um nível de significância de 0.05. Calcule o erro percentual associado. Faça este cálculo utilizando:

- o intervalo de confiança para grandes amostras
- o intervalo de confiança para pequenas amostras

Qual lhe parece mais adequado? Justifique.

c) Utilizando os resultados obtidos por cada aluno faça gráficos equivalentes aos descritos no exercício **Ex5.1-c)** da amostragem casual simples. Compare os resultados com os então obtidos e procure interpretar.

d) Calcule a grandeza da amostra necessária para estimar o volume ha^{-1} da população I com um erro percentual de 5%, a um nível de significância de 0.05, utilizando os valores obtidos em **b)** como se fossem o resultado de um inventário anterior e utilizando dois processos de distribuição da amostra pelos estratos:

- amostragem óptima
- amostragem proporcional à dimensão dos estratos

Comente os resultados

Ex6.11 O mesmo que **Ex6.10**, mas em relação à população II. Compare os resultados e procure interpretar as diferenças encontradas.

Ex6.12 Amostragem da população I estratificada (população IA).

a) Obtenha uma amostra da população IA com 16 indivíduos por estrato e utilize-a para estimar o volume ha^{-1} da população I, a um nível de significância de 0.05. Calcule o erro percentual associado a esta estimativa.

b) Repita o exercício anterior, mas utilizando uma mostra de 5 indivíduos por estrato.

c) Utilizando os resultados obtidos por cada aluno faça gráficos equivalentes aos descritos no exercício **Ex5.1-c)** da amostragem casual simples. Compare os resultados com os então obtidos e com os obtidos nos exercícios **10-c)** e **11-c).** Procure interpretar as diferenças encontradas.

Ex6.13 Cálculo da grandeza da amostra na amostragem estratificada (amostragem óptima)

a) Calcule a dimensão da amostra da população IA que permita estimar o volume ha^{-1} com um erro percentual de 5%, a um nível de significância de 0.05 utilizando as estimativas para a média e variância obtidas com uma amostra de 5 parcelas por estrato.

b) O mesmo que a), mas utilizando as estimativas para a média e variância obtidas com uma amostra de 16 parcelas por estrato.

c) Calcule a dimensão da amostra da população IA que permita estimar o volume ha^{-1} com um erro percentual de 15%, a um nível de significância de 0.05.

d) O mesmo que a), mas utilizando as estimativas para a média e variância obtidas com uma amostra de 16 parcelas por estrato.

e) Compare os resultados obtidos nas diversas alíneas com os obtidos em **10-d)** e **11-d)** e interprete.

6 Amostragem por grupos

Na amostragem por grupos, a unidade de amostragem é constituída por um grupo de indivíduos ("cluster"). No caso do inventário florestal, baseado em parcelas de amostragem, um grupo é constituído por parcelas ou unidades mais pequenas designadas por parcelas ou unidades elementares.

Do ponto de vista prático, a amostragem por grupos apresenta algumas vantagens, em relação a uma amostragem casual simples de igual dimensão:

- em muitas aplicações não existe uma listagem actualizada das unidades elementares que constituem a população, tornando-se mais fácil elaborar esta listagem apenas para os grupos que vierem a fazer parte da amostra.
- do ponto de vista económico é geralmente mais barata na medida em que se evitam em grande parte as despesas de localização das unidades elementares e de deslocação entre estas unidades.

Nos problemas de amostragem apresentados

Na amostragem por grupos há que distinguir dois casos:

- cada grupo (unidade de amostragem) contém o mesmo número de unidades elementares
- número de unidades elementares por grupo é variável e eventualmente desconhecido.

Neste curso, limitar-nos-emos à apresentação do primeiro caso.

6.1 Atributos quantitativos

Sejam:

M -- número de unidades elementares (indivíduos) por grupo

N -- número de grupos (unidades de amostragem)

NM -- número de unidades elementares na população

X_{ij} --valor da variável X no indivíduo j do grupo i

$\mu_i = \frac{1}{M}\sum_{j=1}^{M} x_{ij}$ -- média do grupo i

$\sigma_i^2 = \frac{1}{M}\sum_{j=1}^{M}(x_{ij} - \mu_i)^2$ -- variância dentro do grupo *i*

$\mathrm{T}_i = \sum_{j=1}^{M} x_{ij}$ -- total do grupo *i* $(\mathrm{T}_i = M\ \mu_i)$

$\mu = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} x_{ij}$ -- média da população $\left(\mu = \frac{1}{N}\sum_{i=1}^{N}\mu_i\right)$

$\mu_{\mathrm{T}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{T}_i$ -- média dos totais dos grupos $(\mu_{\mathrm{T}} = M\ \mu)$

$\sigma_{eTg}^2 = \frac{1}{N}N\sum_{i=1}^{\mathrm{N}}(\mathrm{T}_i - \mu_{\mathrm{T}})^2 = \frac{M^2}{N}\sum_{i=1}^{N}(\mu_i - \mu)^2$ -- variância entre os totais dos grupos

6.1.1 Análise de variância

Tal como na amostragem estratificada, também neste caso é possível fazer uma análise de variância

$$x_{ij} - \mu = (x_{ij} - \mu_i) + (\mu_i - \mu)$$

$$(x_{ij} - \mu)^2 = (x_{ij} - \mu_i)^2 + 2(x_{ij} - \mu_i)\ (\mu_i - \mu) + (\mu_i - \mu)^2$$

$$\underbrace{\sum_{i=1}^{N}\sum_{j=1}^{M}(x_{ij} - \mu)^2}_{\text{variação total}} = \underbrace{\sum_{i=1}^{N}\sum_{j=1}^{M}(x_{ij} - \mu_i)^2}_{\substack{\text{variação dentro}\\ \text{dos grupos}}} + \underbrace{M\sum_{i=1}^{N}(\mu_i - \mu)^2}_{\substack{\text{variação entre}\\ \text{os grupos}}}$$

Tabela 9. Análise de variância correspondente a uma amostragem por grupos

Origem da variação	g.l.	Somas de quadrados	Quadrados médios
Dentro dos grupos	$N\,(M\text{-}1)$	$\sum_{i=1}^{N}\sum_{j=1}^{M}\left(x_{ij}-\mu_j\right)^2$	S_{dg}^2
Entre os grupos	$N\text{-}1$	$M\sum_{i=1}^{N}\left(\mu_j-\mu\right)^2$	S_{eg}^2
Total	$N\,M\text{-}1$	$\sum_{i=1}^{N}\sum_{j=1}^{M}\left(x_{ij}-\mu\right)^2$	$S^2=\frac{(N-1)S_{eg}^2+N(M-1)S_{dg}^2}{NM-1}$

6.1.2 Intervalos de confiança para os grupos

Na prática, a amostragem por grupos é uma amostragem casual simples de *n* grupos, estando o erro de amostragem associado apenas à variação entre os grupos uma vez que são observadas todas as unidades elementares dentro de cada grupo pertencente à amostra. Podem então obter-se facilmente os intervalos de confiança a utilizar para uma amostragem por grupos de igual dimensão:

Sejam:

n -- número de grupos que fazem parte da amostra

$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M}x_{ij}$ - média empírica da amostra

$\overline{\mathrm{T}}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{T}_i=M\overline{X}$ -- média empírica dos totais dos grupos

$s_{eg}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{T}_i - \overline{\mathrm{T}})^2$ -- variância empírica entre os totais dos grupos, estimador centrado de σ_{eg}^2

6.1.2.1 Populações normais, pequenas amostras

$$P\ (\overline{\mathrm{T}} - \frac{s_{eg}}{\sqrt{n}} t_{\alpha/2} < \mu_{\mathrm{T}} < \overline{\mathrm{T}} + \frac{s_{eg}}{\sqrt{n}} t_{\alpha/2}) = 1-\alpha$$

6.1.2.2 Populações normais, grandes amostras

$$P\ (\overline{\mathrm{T}} - \frac{s_{eg}}{\sqrt{n}} \frac{N-n}{N} 2_{\alpha/2} < \mu_{\mathrm{T}} < \overline{\mathrm{T}} + \frac{s_{eg}}{\sqrt{n}} \frac{N-n}{N} z_{\alpha/2}) = 1-\alpha$$

6.1.2.3 Populações não normais, grandes amostras

$$P\left(\overline{\mathrm{T}} - \frac{s_{eg}}{\sqrt{n}}\ \frac{N-n}{N}\ z_{\alpha/2} < \mu_{\mathrm{T}} < \overline{\mathrm{T}} + \frac{s_{eg}}{\sqrt{n}}\ \frac{N-n}{N}\ z_{\alpha/2}\right) = 1-\alpha$$

6.1.3 Intervalos de confiança para as unidades elementares

É evidente, que qualquer destes intervalos de confiança se referem aos grupos, sendo, no entanto, extremamente fácil referi-los à unidade elementar. Por exemplo o intervalo c) é equivalente a:

$$P\left(\overline{X} - M\frac{s_{eg}}{M\sqrt{n}}\ \frac{N-n}{N}\ z_{\alpha/2} < \mu < \overline{X} + \frac{s_{eg}}{M\sqrt{n}}\ \frac{N-n}{N}\ z_{\alpha/2}\right) = 1-\alpha$$

6.1.4 Comparação de uma amostragem por grupos com uma amostragem casual simples de igual dimensão

A comparação, em termos de precisão, de uma amostragem por grupos com uma amostragem causal simples de igual dimensão, isto é, com igual número de unidades elementares, pode ser feita por comparação das variâncias da média respectivas.

Amostragem por grupos

$$\text{var}(\overline{X}) = \text{var}\left(\frac{\overline{T}}{M}\right) = \frac{1}{M^2} \quad \text{var}(T) = \frac{1}{M^2}\frac{\alpha_{eg}^2}{n}\frac{N-n}{N-1}$$

Mas $\alpha_{eg}^2 = \frac{N-1}{N} \; M \; S_{eg}^2$

pelo

$$\text{var}(\overline{X}) = \frac{1}{M}\frac{S_{eg}^2}{n}\frac{N-n}{N}$$

Amostragem casual simples de dimensão nM

$$\text{var}(\overline{X}) = \frac{\sigma^2}{nM}\frac{NM-nM}{NM-1}$$

$$= \frac{S^2}{nM}\frac{NM-nM}{NM} = \frac{S^2}{NM}\frac{N-n}{N}$$

Uma amostragem por grupos será então mais precisa que uma amostragem casual simples de igual dimensão (n>30), desde que

$$\frac{1}{M}\frac{S_{eg}^2}{n}\frac{N-n}{N} < \frac{S^2}{nM} \quad \frac{N-n}{N}$$

$$\downarrow$$

$$S_{eg}^2 < S^2$$

ou seja, desde que o quadrado médio entre grupos seja inferior ao quadrado médio da população total.

Atendendo a que

$$S_{eg}^2 = \frac{NM-1}{N-1}S^2 - \frac{N\ (M-1)}{N-1}S_{dg}^2$$

pode concluir-se também que uma amostragem por grupos é mais eficiente que uma amostragem casual simples de igual dimensão se

$$S_{dg}^2 > S^2$$

Conclui-se então que para que uma amostragem por grupos seja eficiente os grupos deverão ser muito semelhantes entre si, embora internamente muito variáveis. Infelizmente, na prática, esta condição é frequentemente difícil de satisfazer uma vez que os grupos são geralmente constituídos por unidades elementares próximas no espaço as quais são frequentemente semelhantes.

6.2 Análise da amostragem sistemática como um caso particular da amostragem por grupos

6.2.1 Análise de variância

Note-se que uma amostragem sistemática pode ser interpretada como uma amostragem por grupos em que se observa apenas um dos *k* grupos que constituem a população.

Sejam:

$X\left(\mu,\sigma^2\right)$ -- variável em estudo

N – dimensão da população (admitamos que *N* é múltiplo de *n*)

$\left(X_1,X_2,...,X_n\right)$ -- amostra sistemática de dimensão *n*

x_{ij} -- valor da variável *X* no indivíduo *j* da amostra sistemática *i*

$\mu_i = \overline{X}_{si} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$ -- média da amostra sistemática *i*

Tal como na amostragem por grupos, de que a amostragem sistemática é um caso particular, pode efectuar-se a seguinte análise de variância (tabela 10).

Tabela 10. Análise de variância correspondente a uma amostragem sistemática

Origem da variação	g.l.	Somas de quadrados	Quadrados médios
Dentro das amostras	$k\,(n\text{-}1)$	$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\mu_i)^2$	S^2_{da}
Entre amostras	k-1	$n\sum_{i=1}^{k}(\mu_i-\mu)^2$	S^2_{ea}
Total	$k\,n$-1	$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\mu)^2$	$S^2 = \frac{k\ (n-1)S^2_{da} + (k-1)S^2_{ea}}{k\ n-1}$

A variância da média sistemática $\overline{X}_{si}\,(=\mu_i)$ é, por definição:

$$\mathrm{var}(\overline{X}_{si}) = \frac{1}{k}\sum_{i=1}^{k}(\mu_i-\mu)^2 = \frac{S^2_{ea}(k-1)}{n\ k} = \frac{S^2_{es}(k-1)}{N}$$

Da análise da variância, tira-se que

$$(N-1)\ S^2 = (k-1)\ S^2_{ea} + k\ (n-1)\ S^2_{da}$$

$$(N-1)\ S^2 = N\ \text{var}\left(\overline{X}_{si}\right) + k\ (n-1)\ S_{da}^2$$

$$\text{var}\left(\overline{X}_{si}\right) = \frac{N-1}{N}\ S^2 - \frac{k(n-1)}{N}\ S_{da}^2$$

6.2.2 Comparação de uma amostragem sistemática com uma amostragem casual simples de igual dimensão

Podem agora comparar-se as precisões de uma amostragem sistemática e de uma amostragem casual simples, por comparação das variâncias das respectivas médias $\overline{X}_{si}$ e $\overline{X}$. Uma amostragem sistemática é mais precisa que uma amostragem casual simples de igual dimensão se

$$\text{var}\left(\overline{X}_{si}\right) < \text{var}\left(\overline{X}\right)$$

$$\frac{N-1}{N}\ S^2 - \frac{k(n-1)}{N}\ S_{da}^2 < \frac{N-n}{N}\ \frac{S^2}{n}$$

$$\frac{k\ (n-1)}{N}\ S_{da}^2 > \left(\frac{N-1}{N} - \frac{N-n}{N\ n}\right)\ S^2$$

$$k\ (n-1)\ S_{da}^2 > (N-1-k+1)\ S^2$$

$$k\ (n-1)\ S_{da}^2 > k\ (n-1)\ S^2$$

$$S_{da}^2 > S^2$$

Este resultado, aplicável à amostragem por grupos em geral, implica que a amostragem sistemática é mais precisa que uma amostragem casual simples de igual dimensão, se o quadrado médio dentro das amostras for maior que o quadrado médio da população.

6.3 Exercícios

Ex7.1 Considere a população I dividida em 100 grupos de 4 indivíduos contíguos.

a) Obtenha uma amostra de 12 desses grupos, correspondente portanto a 48 indivíduos (amostragem por grupos)

b) Utilize a amostra obtida em **a)** para estimar o volume ha^{-1} da população I a um nível de significância de 0.05. Calcule o erro percentual associado a esta estimativa.

c) Compare com os resultados obtidos com a amostragem casual simples e com os diferentes esquemas de amostragem estratificada considerados. Procure interpretar as diferenças encontradas.

d) Utilizando os resultados obtidos por cada aluno faça gráficos equivalentes aos descritos no exercício **Ex5.1-c)** da amostragem casual simples. Compare os resultados com os então obtidos e com os obtidos nos exercícios da amostragem estratificada**.** Procure interpretar as diferenças encontradas.

Ex7.2 Considere agora a população I dividida em 100 grupos constituidos por 4 indivíduos não contíguos no espaço. Resolva novamente o exercício **1**, procurando interpretar as diferenças encontradas.

Ex7.3 Resolva agora os exercícios 1 e 2, mas em relação à população II. Interprete as diferenças encontradas.

7 Referências bibliográficas

Borralho, R., F. Rego, F. Palomares e A. Hora, 1995. The distribution of the Egyptian Mongoose *Herpestes ichneumon* (L.) in Portugal. *Mammal Rev*. 25(4):229-236.

Cochran, W. G., 1977. *Sampling techniques*, 3ª ed. Wiley series in probability and mathematical statistics. John Wiley & Sons, New York.

Cumbre, M. F., 1999. Avaliação da qualidade tecnológica de pranchas de cortiça por amostragem em pilha. Relatório do Trabalho de Fim do Curso de Engenharia Florestal, Instituto Superior de Agronomia, Lisboa, Portugal.

Loetsch, F. e K. E. Haller, 1973. *Forest inventory*, vol. I. BLV Verlagsgesellschaft Müchen, Bern, Wien. (trad. para inglês de E. F. Brunig).

Scheaffer, R. L., W. Mendenhall e L. Off, 1990. Elementary survey sampling. PWS-KENT, Boston (cit. in Shiver e Borders, 1998).

Shiver, B. D. e B. E. Borders, 1998. Sampling techniques for forest resource inventory. John Wiley & Sons, New York.

Schumacher, F. X. e R. A. Chapman, 1954. *Sampling methods in forestry and range management*. Duke University. Durham.

Tomé, M., T. Oliveira e J. C. Paul, 1992. *Inventário florestal. Concelho de Oliveira do Hospital*. Câmara Municipal de Oliveira do Hospital.

FORMULÁRIO PARA APOIO À DISCIPLINA DE INVENTÁRIO FLORESTAL

Amostragem aplicada ao Inventário Florestal

Estatísticas correspondentes aos parâmetros mais importantes de uma população

Parâmetro[1]	Estatística correspondente
Média $\mu_1' = \mu = E(X) = \frac{1}{N}\sum_{i=1}^{N} x_i$	**Média empírica** $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Momento de ordem k $\mu_k' = E\left(X^k\right) = \frac{1}{N}\sum_{i=1}^{N} x_i^k$	**Momento empírico de ordem k** $m_k' = \frac{1}{n}\sum_{i=1}^{n} x_i^k$
Variância $\mu_2 = \sigma^2 = E(X-\mu)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$	**Variância empírica** $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_1-\overline{x}.)^2$
Momento central de ordem k $\mu_2^k = E(X-\mu)^k = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^k$	**Momento central empírico de ordem k** $m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^k$

[1] As expressões sob a forma de somatórios correspondem a populações finitas

Estimadores da média, variância e variância da média nas amostragens som e sem reposição

	Amostragem com reposição	Amostragem sem reposição
Estimador centrado de μ	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
Estimador centrado de σ^2	$s_c^2 = \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n-1}$	$s_{s,c}^2 = s_c^2\ \frac{N-1}{N}$
Variância de $\overline{X}$ - $\sigma_{\overline{X}}^2$	$\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}$	$\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}\frac{N-n}{N-1}$
Estimador centrado de $\sigma_{\overline{X}}^2$	$s_{\overline{X}}^2 = \frac{s_c^2}{n}$	$s_{\overline{X}}^2 = \frac{s_c^2}{n}\ \frac{N-n}{N}$

1 Amostragem simples, atributos quantitativos

1.1 Intervalos de confiança

1.1.1 Populações normais e não normais, grandes amostras (n>30)

$$P\left(\overline{X}_n - s_{\overline{X}}\ z_{\alpha/2} \le \mu \le \overline{X}_n + s_{\overline{X}}\ z_{\alpha/2}\right) = 1-\alpha$$

$s^2_{\overline{X}} = \frac{s_c^2}{n}\ \frac{N-n}{N}$, na amostragem sem reposição

Erro absoluto: $E = s_{\overline{X}}\ z_{\alpha/2}$

Erro percentual: $E\% = \frac{E}{\overline{X}}\ 100$

1.1.2 Populações normais, pequenas amostras

$$P\left(\overline{X}_n - t_{\alpha/2}\frac{s_c}{\sqrt{n}} \le \mu \le \overline{X}_n + t_{\alpha/2}\frac{s_c}{\sqrt{n}}\right) = 1-\alpha$$

Erro absoluto: $E = \frac{s_c}{\sqrt{n}}\ t_{\alpha/2}$

Erro percentual: $E\% = \frac{E}{\overline{X}}\ 100$

1.2 Cálculo da grandeza da amostra

$$E = \frac{E\%\ \overline{X}}{100}$$

1.2.1 Populações normais e não normais, grandes amostras (n>30)

$$n = \frac{s_c^2 \; z^2_{\alpha/2}}{E^2 + \frac{z^2_{\alpha/2} s_c^2}{N}} = \frac{s_c^2 \; z^2_{\alpha/2}}{\left(\frac{E\% \; \overline{X}}{100}\right)^2 + \frac{z^2_{\alpha/2} \; s_c^2}{N}}$$

1.2.2 Populações normais, pequenas amostras

$$n = \frac{t^2_{\alpha/2} s_c^2}{E^2} = \frac{t^2_{\alpha/2} s_c^2}{\left(\frac{E\% \; \overline{X}}{100}\right)^2}$$

Não esquecer que, neste caso, o cálculo é feito de modo iterativo até que haja acordo entre o valor de n e o número de graus de liberdade utilizados para o t-Student

2 Amostragem simples, atributos qualitativos

2.1 Intervalos de confiança

2.1.1 Populações normais e não normais, grandes amostras (n p > 10)

$$P\left(p^* - z_{\alpha/2}\sqrt{\frac{p^* q^*}{n-1}} \le p \le p^* + z_{\alpha/2}\sqrt{\frac{p^* q^*}{n-1}}\right) \approx 1 - \alpha$$

Erro absoluto: $E = \sqrt{\frac{p^* q^*}{n-1}} \; z_{\alpha/2}$

Erro percentual: $E\% = \frac{E}{p^*} \; 100$

2.2. Cálculo da grandeza da amostra

$$E = \frac{E\%\ \ p^*}{100}$$

2.2.1 Populações normais e não normais, grandes amostras (n p > 10)

$$n = \frac{z^2_{\alpha/2}\ \ p^* q^*}{E^2} + 1 \ \approx\ \frac{z^2_{\alpha/2}\ \ p^* q^*}{E^2} \ =\ \frac{z^2_{\alpha/2}\ \ p^* q^*}{\left(\frac{E\%\ \ p^*}{100}\right)^2}$$

3 Amostragem estratificada, atributos quantitativos

3.1 Intervalos de confiança

3.1.1 Populações normais e não normais, grandes amostras (n>30, nj > 10)

$$P\left(\overline{X}_{st} - z_{\alpha/2} s_{\overline{X}_{st}} \le \mu \le \overline{X}_{st} + z_{\alpha/2} s_{\overline{X}_{st}}\right) = 1 - \alpha$$

$$\overline{X}_{st} = \sum_{j=1}^{M} P_j \overline{X}_j$$

$$s^2_{\overline{X}_{st}} = \sum_{j=1}^{M} P_j^2 \ \ \frac{s^2_{j,c}}{n_j} \ \ \frac{N_j - n_j}{N_j}$$

Erro absoluto: $E = s_{\overline{X}_{st}}\ \ z_{\alpha/2}$

Erro percentual: $E\% = \frac{E}{\overline{X}_{st}}\ \ 100$

3.1.2 Populações normais, pequenas amostras

$$P\left(\overline{X}_{st} - t_{\alpha/2}\frac{s_{st,c}}{\sqrt{n}} \le \mu \le \overline{X}_{st} + t_{\alpha/2}\frac{s_{st,c}}{\sqrt{n}}\right) = 1-\alpha$$

$$\overline{X}_{st} = \sum_{j=1}^{M} P_j \overline{X}_j$$

$$s_{st,c}^2 = \underbrace{\frac{1}{n}\sum_{j=1}^{M}\sum_{i=1}^{n_j}\left(x_{ij} - \overline{X}_j\right)^2}_{s_{st}^2} \cdot \frac{n}{n-M}$$

Erro absoluto: $E = \frac{s_{st,c}}{\sqrt{n}}\ t_{\alpha/2}$

Erro percentual: $E\% = \frac{E}{\overline{X}_{st}}\ 100$

3.2 Cálculo da grandeza da amostra

$$E = \frac{E\%\ \overline{X}_{st}}{100}$$

3.2.1 Populações normais e não normais, grandes amostras (n>30)

3.2.1.1 Amostragem proporcional à dimensão dos estratos

Cálculo da dimensão da amostra (n):

$$n = \frac{z_{\alpha/2}^2 \sum_{j=1}^{M} P_j s_{j,c}^2}{E^2 + \frac{z_{\alpha/2}^2 \sum_{j=1}^{M} P_j s_{j,c}^2}{N}} = \frac{z_{\alpha/2}^2 \sum_{j=1}^{M} P_j s_{j,c}^2}{\left(\frac{E\%\ \overline{X}_{st}}{100}\right)^2 + \frac{z_{\alpha/2}^2 \sum_{j=1}^{M} P_j s_{j,c}^2}{N}}$$

Distribuição da amostra pelos diferentes estratos (cálculo dos nj): $n_j = P_j \ .n$

3.2.1.2 Amostragem óptima

Cálculo da dimensão da amostra (n):

$$n = \frac{\left(\sum_{j=1}^{M} P_j s_{j,c}\right)^2 z_{\alpha/2}^2}{E^2 + \frac{z_{\alpha/2}^2 \sum_{j=1}^{M} P_j s_{j,c}^2}{N}} = \frac{\left(\sum_{j=1}^{M} P_j s_{j,c}\right)^2 z_{\alpha/2}^2}{\left(\frac{E\% \ \overline{X}_{st}}{100}\right)^2 + \frac{z_{\alpha/2}^2 \sum_{j=1}^{M} P_j s_{j,c}^2}{N}}$$

Distribuição da amostra pelos diferentes estratos (cálculo dos nj):

$$n_j = \frac{P_j s_{j,c}}{\sum_{j=1}^{M} P_j s_{j,c}} \quad \Leftrightarrow \quad n = \frac{N_j s_{j,c}}{\sum_{j=1}^{M} N_j s_{j,c}} n$$

3.2.2 Populações normais, pequenas amostras (proporcional à dimensão dos estratos)

Cálculo da dimensão da amostra (n):

$$n = \frac{t_{\alpha/2}^2 \ s_{st,c}^2}{E^2} = \frac{t_{\alpha/2}^2 \ s_{st,c}^2}{\left(\frac{E\% \ \overline{X}_{st}}{100}\right)^2}$$

Não esquecer que, neste caso, o cálculo é feito de modo iterativo até que haja acordo entre o valor de n e o número de graus de liberdade utilizados para o t-Student

Distribuição da amostra pelos diferentes estratos (cálculo dos nj): $n_j = P_j \ .n$