



N.º 29 – O QUE É UMA SONDAGEM? COMO É TRANSMITIDO O RESULTADO DE UMA SONDAGEM? O QUE É UM INTERVALO DE CONFIANÇA?

Por: Maria Eugénia Graça Martins
Departamento de Estatística e Investigação Operacional da FCUL
memartins@fc.ul.pt

INTRODUÇÃO

O objetivo de uma sondagem é o de recolher informação acerca de uma *população*, selecionando e observando um conjunto de elementos dessa *população*.



SONDAGEM – Estudo estatístico de uma *população*, feito através de uma *amostra*, destinado a estudar uma ou mais características tais como elas se apresentam nessa população¹.

Considere-se a população constituída por todos os potenciais eleitores para as próximas eleições legislativas. De um modo geral e com alguma antecedência relativamente à data das eleições, os partidos políticos pretendem obter uma estimativa de como se fará a distribuição de votos ou obter outras características desta população. O tempo, custos e outros inconvenientes impedem os partidos de fazer a pergunta a todos os elementos da população, pelo que a informação pretendida será obtida inquirindo apenas uma parte do conjunto de todos os eleitores – (*amostra*), mas com o objetivo de tirar conclusões para o conjunto todo (*população*).

Às características numéricas da população para a qual se pretende obter informação damos o nome de **parâmetros**. Assim, relativamente à população constituída pelos

¹ Por vezes, confunde-se sondagem com amostragem. No entanto, a amostragem diz respeito ao procedimento da recolha da amostra qualquer que seja o estudo estatístico que se pretenda fazer, pelo que a amostragem é uma das fases das sondagens, já que estas devem incluir ainda o estudo dos dados recolhidos, assim como a elaboração do relatório.

potenciais eleitores das legislativas, alguns parâmetros que pode ter interesse conhecer são:

- Idade média dos potenciais eleitores;
- Percentagem de eleitores que estão decididos a votar;
- Percentagem de eleitores que estão decididos a votar em cada partido;
- Etc.

Os parâmetros são estimados por **estatísticas**, números que se calculam a partir dos valores obtidos como resultado da observação da variável de interesse nos elementos selecionados para a amostra (vamos também designar por **amostra** o conjunto destas observações ou dados). Como, de um modo geral, podemos obter muitas amostras diferentes, embora da mesma dimensão, teremos muitas estimativas do(s) parâmetro(s) em estudo. Tantas as amostras diferentes que se puderem selecionar da população (2 amostras da mesma dimensão serão diferentes, se diferirem pelo menos num dos elementos selecionados), tantas as estimativas, eventualmente diferentes, que se podem calcular para o parâmetro. Podemos considerar que todas estas estimativas são os valores observados de uma função dos elementos da amostra a que se dá o nome de **estimador**.

Assim:

Um **parâmetro** é uma característica numérica da **população**, enquanto a **estatística** é uma característica numérica da **amostra**. Um **estimador** é uma função dos elementos da amostra, que se utiliza para estimar parâmetros. Ao valor do **estimador** calculado para uma determinada amostra, dá-se o nome de **estimativa** (ou **estatística**).

INTERVALO DE CONFIANÇA PARA O VALOR MÉDIO OU MÉDIA POPULACIONAL

Admitamos que o **parâmetro** a estudar é a média (populacional) das idades de todos os potenciais eleitores. Para obter uma estimativa deste valor, recolhe-se uma amostra de potenciais eleitores, regista-se a idade de cada um e calcula-se a média das idades obtidas. Por exemplo, suponha-se que se recolheu uma amostra de 15 eleitores e os dados obtidos (registos das idades dos 15 eleitores) foram

54 29 92 33 81 57 41 60 20 42 37 36 57 26 72

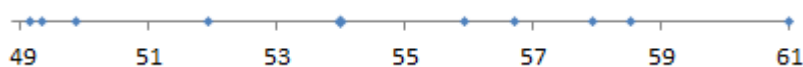
A média da amostra anterior é 49,1 anos, de modo que este valor é uma estimativa para o valor médio ou média (populacional) das idades de toda a população.

Será que nos podemos dar por satisfeitos? De maneira nenhuma! Se recolhermos várias amostras da mesma dimensão, o resultado obtido pode ser diferente de amostra para amostra, isto é, as várias médias calculadas, uma para cada uma das amostras, serão eventualmente diferentes.

Por exemplo, suponhamos que se recolheram 10 amostras de dimensão 15, tendo-se registado as seguintes idades:

Amostras									
1	2	3	4	5	6	7	8	9	10
54	54	64	87	58	64	26	75	50	70
80	29	52	70	48	32	67	80	68	68
32	92	50	69	51	18	40	40	33	49
47	33	71	49	35	25	42	54	35	66
62	81	50	64	58	79	87	48	61	57
42	57	54	32	55	37	38	81	57	47
83	41	32	84	77	41	78	57	46	34
54	60	73	22	58	41	79	46	54	55
39	20	47	71	36	44	63	59	83	49
49	42	42	67	71	39	74	36	53	26
36	37	49	48	70	76	73	43	51	58
39	36	57	51	56	60	20	59	31	43
80	57	44	60	47	90	21	74	78	95
89	26	29	86	61	24	50	84	48	68
83	72	26	55	70	78	21	42	62	54

Para cada uma das amostras anteriores calculou-se a média, tendo-se obtido os valores registados no seguinte gráfico de pontos:



Obtivemos 10 estimativas variando entre 49,1 e 61,0. Qual destas estimativas é a melhor? Qual é a que está mais perto da média das idades de todos os eleitores? Não sabemos, nem temos processo de saber, já que a média das idades de todos os eleitores é desconhecida e por isso é que estamos a estimá-la!

Não desanimemos! Vamos tentar resolver a situação, abordando o problema de outra forma.

Mas, primeiro, vejamos o que acontece se aumentar a dimensão das amostras recolhidas. Em vez de amostras de dimensão 15, vamos considerar, por exemplo, amostras de dimensão 100. Considerámos, então, 10 amostras de dimensão 100 e calculámos as médias, cujos valores são apresentados no gráfico seguinte:



Qual a diferença entre os dois gráficos? As estimativas obtidas com amostras de maior dimensão estão mais perto umas das outras, variam entre 52,2 e 58,4 e assim esperamos que estejam mais perto do valor do parâmetro (desconhecido!). Nesta altura é necessário fazer uma chamada de atenção muito importante: estamos a partir do princípio de que as amostras foram “bem” selecionadas² e são representativas da população de onde foram recolhidas.

Ao estimar o parâmetro “valor médio” ou média (populacional) das idades de todos os potenciais eleitores, estamos a utilizar o *estimador Média* (amostral). Mesmo que as diferentes amostras tenham a mesma dimensão, as estimativas fornecidas por este estimador são diferentes de amostra para amostra; considerámos 10 amostras e obtivemos 10 valores para o estimador, ou seja, 10 estimativas. Assim, a questão para a qual gostaríamos de ter resposta, é a seguinte:

² Consultar explicação mais detalhada sobre seleção de amostras no Curso de Introdução à Inferência Estatística do ALEA http://www.alea.pt/html/statofic/html/dossier/doc/Modulo1-Int_AmostragemFinal.pdf

Como se comportam, relativamente ao parâmetro em estudo, todas as estimativas fornecidas por um dado estimador, para todas as amostras possíveis, de uma determinada dimensão? Ou seja, como é que se distribuem todos os valores obtidos pelo estimador para todas as amostras possíveis? Ou, no caso que estamos a tratar, qual a distribuição de amostragem³ do estimador **Média**, que representaremos, daqui em diante, por \bar{X} ?

A resposta à pergunta anterior é crucial, como veremos mais à frente. Para já, podemos adiantar que normalmente não se conhece a forma da distribuição de amostragem exata da **Média** \bar{X} , mas sabe-se que⁴:

➤ Resultado 1

Se a população tiver dimensão grande, valor médio μ e desvio padrão σ , então, para amostras de dimensão n , o valor médio (média de todas as estimativas fornecidas pelo estimador, para todas as amostras possíveis) do estimador \bar{X} é também μ e o seu desvio padrão é $\frac{\sigma}{\sqrt{n}}$.

E quanto à forma da distribuição? Temos o seguinte resultado, que é de grande relevância, na medida em que nos vai resolver o problema da estimação que estamos a tratar:

➤ Resultado 2

Quando se faz amostragem sem reposição e as populações têm dimensão razoavelmente grande ou no caso de a amostragem ser com reposição, as populações terem qualquer dimensão e as amostras têm dimensão grande (é usual considerar maior ou igual a 30), a distribuição de amostragem do estimador **Média** \bar{X} pode ser aproximada pela distribuição Normal, independentemente da distribuição dos valores da variável sobre os elementos da população de onde se selecionam as amostras (ou seja, independentemente da distribuição da população subjacente).

Este resultado é uma consequência de um dos teoremas mais importantes da Probabilidade, o **Teorema Limite Central**, que legitima a grande utilização do modelo **Normal** ou **Gaussiano**⁵.

Repare-se que os resultados 1 e 2 permitem concluir que as estimativas fornecidas pelo estimador **Média** se distribuem de forma aproximadamente simétrica em torno do parâmetro valor médio (μ) que se está a estimar e que, quanto maior for a dimensão das amostras consideradas, menor será a variabilidade (σ/\sqrt{n}) com que esses valores se distribuem em torno do parâmetro.

³ À distribuição de um estimador dá-se o nome de *distribuição de amostragem*.

⁴ Ver http://www.alea.pt/html/statofic/html/dossier/doc/Modulo2-Int_EstimacaoFinal.pdf, página 39 e seguintes.

⁵ Ver http://www.alea.pt/html/statofic/html/dossier/doc/Modulo2-Int_EstimacaoFinal.pdf, página 40 e seguintes.

O comportamento da distribuição de amostragem da **Média** \bar{X} tem consequências muito importantes no que diz respeito à estimação do parâmetro “valor médio” ou média populacional, pelo que vamos aproveitá-lo para encarar este problema (o da estimação do parâmetro) de um outro ângulo. Em vez de procurarmos um valor (*estimativa pontual*) como aproximação do valor do parâmetro desconhecido, neste caso a média da população, vamos procurar obter um intervalo (estimativa intervalar ou *intervalo de confiança*) que, com uma determinada confiança, contenha o valor desse parâmetro!⁶

Vamos então procurar um intervalo aleatório $[A, B]$ que, com uma “grande probabilidade”, por exemplo, 95%, contenha o parâmetro μ :

$$P([A, B] \text{ conter } \mu) = 95\%$$

Ora, é precisamente na construção destes **intervalos de confiança** que vamos aproveitar o facto de a distribuição de amostragem da **Média** \bar{X} poder ser aproximada pelo modelo Normal, com valor médio igual ao valor médio μ (parâmetro que estamos a estimar) da População e desvio padrão igual a σ/\sqrt{n} , onde σ é o desvio padrão da população. Como o desvio padrão da População é quase sempre desconhecido, vamos estimá-lo pelo desvio padrão amostral s , de modo que um valor aproximado para o desvio padrão do estimador \bar{X} , também conhecido como erro padrão, é s/\sqrt{n} .

Então, tendo em consideração as propriedades da distribuição Normal, podemos escrever:

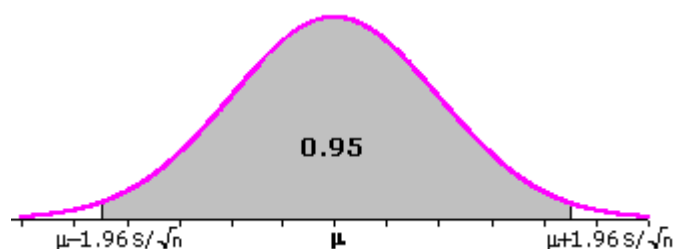
$$P(-1,96 \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq 1,96) \approx 0,95 \quad (1)$$

O valor 1,96 pode ser obtido consultando uma tabela, a calculadora ou a folha de Excel. De (1) vem

$$P(\mu - 1,96 s/\sqrt{n} \leq \bar{X} \leq \mu + 1,96 s/\sqrt{n}) \approx 0,95$$

ou

$$P(\bar{X} - 1,96 s/\sqrt{n} \leq \mu \leq \bar{X} + 1,96 s/\sqrt{n}) \approx 0,95$$



⁶ Ver http://www.alea.pt/html/statofic/html/dossier/doc/Modulo3-Int_InferenciaEstatisticaFinal.pdf

Então a expressão de um **intervalo de confiança** com uma confiança de 95% é dada pelo intervalo

$$[\bar{x} - 1,96 \times s / \sqrt{n} ; \bar{x} + 1,96 \times s / \sqrt{n}]$$

onde \bar{x} e s são, respetivamente, a média e o desvio padrão de uma amostra de dimensão n , recolhida para estimar μ .

A metade da amplitude do intervalo de confiança atribui-se a designação *margem de erro*.

➤ Afinal, o que significa um intervalo de 95% de confiança?

Significa que, se recolhermos muitas amostras de dimensão n , calcularmos as médias e os desvios padrões dessas amostras e construirmos os intervalos de confiança respetivos, utilizando a expressão anterior, cerca de 95% desses intervalos conterão o valor médio μ , enquanto os restantes 5% não o conterão. Não temos a certeza de que um dado intervalo, em particular, contenha o parâmetro desconhecido, mas estamos confiantes de que assim aconteça, isto é, estamos 95% confiantes que o intervalo que calculámos a partir da amostra selecionada (na prática, só seleccionámos uma amostra) contenha o valor do parâmetro μ .

E se pretendermos um intervalo de 90% de confiança? Ou de 99%? A forma geral do intervalo de confiança é

$$[\bar{x} - z \times s / \sqrt{n} ; \bar{x} + z \times s / \sqrt{n}]$$

onde o valor de z depende da confiança com que se quer construir o intervalo. Alguns valores (obtidos a partir da distribuição da Normal (0,1)) são

Confiança	z
90%	1,645
95%	1,960
97,5%	2,326
99%	2,576
99,5%	3,090

Caso prático

Como proceder, então, para obter um intervalo de confiança, com uma confiança de 95% para a idade média dos potenciais eleitores?

Passo 1 - Recolher uma amostra da população dos potenciais eleitores. Repare-se que da expressão do intervalo de confiança se conclui que, quanto maior for a dimensão n da amostra, melhor será a amplitude do intervalo (quanto menor for a amplitude, melhor!). Recolhemos uma amostra de dimensão 40, que apresentamos a seguir:

24 65 33 25 79 73 52 63
18 46 28 97 53 87 26 78
89 62 45 30 57 82 66 52
19 41 75 58 55 42 51 18
43 82 46 36 57 59 93 65

Passo 2 – Calcular a média e o desvio padrão da amostra selecionada. Para a amostra anterior, temos

$$\bar{x} = 54,25 \text{ e } s = 21,99$$

Passo 3 – Obter os limites do intervalo de confiança

$$[47,4; 61,1]$$

Passo 4 – Concluir, dizendo que um intervalo de 95% de confiança para a média das idades da população em estudo é [47,4 anos; 61,1 anos] ou que uma estimativa para a idade média é 54,25 anos, com uma margem de erro de 6,82 anos e uma confiança de 95%.

INTERVALO DE CONFIANÇA PARA A PROPORÇÃO POPULACIONAL

Suponhamos agora que o que se pretendia era estimar a proporção (ou percentagem) de eleitores que pensam votar no partido SOL (fictício). Sendo agora o parâmetro em estudo a proporção populacional, será natural estimar o valor deste parâmetro através da proporção (amostral) de eleitores que, numa amostra recolhida da população de eleitores, pensam votar no partido SOL.

Consideremos então a população de potenciais eleitores e seja **p** a proporção (desconhecida) de eleitores que pensam votar no partido Sol. Repare-se que a proporção **p** não é mais do que uma média (populacional) de 0's e 1's, em que atribuímos o valor 1 a um elemento da população que pertença à categoria em estudo (o que, neste caso, significa votar no partido SOL) e o valor 0 a um elemento que não pertença a essa categoria.

Assim, como a proporção **p** é o valor médio de uma população cujos elementos são 0's e 1's, o estudo anteriormente feito para a estimação do valor médio será facilmente adaptado para a estimação da proporção. Para esta população tão particular, constituída por 0's e 1's, em que a proporção populacional é a média populacional, a **Proporção** amostral também será a **Média** (amostral), que será, assim, o estimador intuitivo para a proporção populacional. Assim, não temos mais do que transportar para o estudo da proporção os resultados obtidos quando se considerou o estimador **Média**.

Temos então uma população constituída por 0's e 1's em que a proporção de 1's é **p** e a proporção de 0's é **(1-p)**:

Classe	Freq. relativa
0	(1-p)
1	p
Total	1

É imediato que o valor médio e a variância (populacional) desta população são, respetivamente:

$$\mu = p \quad (=0 \times (1-p) + 1 \times p) \quad \text{e} \quad \sigma^2 = p(1-p) \quad (=(0-p)^2 \times (1-p) + (1-p)^2 \times p)$$

Representando o estimador da proporção **p** por \hat{p} e adaptando os resultados obtidos para o estimador **Média** \bar{X} , temos o seguinte resultado.

➤ Resultado

Suponhamos que se seleciona uma amostra aleatória simples de uma População de dimensão grande, ou que se seleciona uma amostra aleatória, com reposição de uma população de dimensão qualquer, em que a característica em estudo está presente numa proporção p (desconhecida). Então, se a dimensão n da amostra for suficientemente grande (um valor que é usual considerar como suficientemente grande é 30), a distribuição de amostragem da **Proporção** amostral \hat{p} pode ser aproximada pela distribuição Normal com valor médio p e desvio padrão $\sqrt{\frac{p(1-p)}{n}}$.

Assim, a expressão de um intervalo de 95% de confiança para a proporção p tem a seguinte forma:

$$[\hat{p} - 1,96 \sqrt{\frac{p(1-p)}{n}} ; \hat{p} + 1,96 \sqrt{\frac{p(1-p)}{n}}]$$

Como p é desconhecido, é substituído por uma sua estimativa \hat{p} , pelo que a forma de um intervalo de confiança para a proporção tem o seguinte aspeto:

$$[\hat{p} - 1,96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} ; \hat{p} + 1,96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

Caso prático

Como proceder, então, para obter um intervalo de confiança, com uma confiança de 95%, para a proporção dos eleitores que pensam votar no partido SOL?

Passo 1 - Recolher uma amostra da população dos potenciais eleitores. Decidimos recolher uma amostra de dimensão 50 e os dados obtidos foram os seguintes, (representou-se por 1 uma resposta de um eleitor que pensa votar no SOL):

0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0
0 0 0 0 0 1 0 0 0 0 1 0 1 1 1 0 1 0 1 1 0 0 1 1 0

Passo 2 - Calcular a proporção (amostral) de eleitores que pensam votar no SOL (frequência relativa de 1's. Para a amostra anterior, temos

$$\hat{p} = 0,28$$

Passo 3 - Obter os limites do intervalo de confiança para a proporção p

$$[0,156; 0,404]$$

Passo 4 - Concluir, dizendo que um intervalo de 95% de confiança para a proporção de eleitores da população em estudo, que pensam votar no partido Sol, é [15,6%; 40,4%], ou então dizer que uma estimativa para a proporção de eleitores que pensam votar no SOL é de 28%, com uma margem de erro de 12,4% e uma confiança de 95%.

Nota – É possível obter um intervalo de confiança com uma determinada confiança e com uma margem de erro inferior a determinado valor d , fixado antes da recolha da amostra. Neste caso, a dimensão da amostra necessária ficará condicionada por esta escolha

(Consultar o Curso de Introdução à Inferência Estatística do ALEA,

http://www.alea.pt/html/statofic/html/dossier/doc/Modulo3-Int_InferenciaEstatisticaFinal.pdf, página 75).