Unidade IV
Inferência Estatística

4.6. Dimensionamento de amostras

Um dos problemas mais comuns relacionados à inferência estatística é a determinação do tamanho de amostra suficiente para representar uma população. A obtenção do tamanho da amostra está relacionada com dois aspectos básicos: (a) qual é o objetivo final do processo de amostragem e com que precisão se deseja alcançá-lo e (b) qual a variabilidade (aproximada ou estimada) da variável em estudo. Abordaremos a seguir um processo para dimensionar uma amostra que tem como objetivo a construção de um intervalo de confiança para a média da população.

Já foi discutido anteriormente que amostragem é o processo de obtenção das amostras e que, dependendo deste processo, as amostras podem ser probabilísticas ou não probabilísticas. Devemos salientar, no entanto, que o processo para dimensionar amostras que será tratado aqui se aplica apenas a um tipo especial de amostra probabilística, denominada *amostra aleatória simples* ou *amostra casual simples*.

Amostra casual simples é aquela obtida de modo que todos os elementos da população tenham a mesma probabilidade de fazer parte da amostra, ou ainda, que todas as possíveis amostras tenham igual probabilidade de ser selecionada. Uma vez que todas as possíveis amostras têm a mesma chance de ocorrer, este processo de amostragem é recomendado quando os elementos da população não apresentam grande variabilidade para a característica que se pretende analisar.

Ao dimensionarmos uma amostra, necessitamos do conhecimento prévio da variabilidade da população e do grau de precisão (γ) desejado para a amostragem. O grau de precisão é estabelecido de acordo com os objetivos da amostragem e como, geralmente, a variância da população de onde a amostra será retirada é um valor desconhecido, podemos utilizar uma amostra piloto, denominada pré-amostra, com o objetivo de estimar a variância populacional. A partir dessas informações é possível dimensionar o tamanho de amostra suficiente para estimar, por intervalo, a média da população.

Quando as populações são consideradas infinitas ou as amostras são retiradas *com reposição*, sabe-se que a estimativa do erro padrão da média é

$$S(\bar{X}) = \frac{S}{\sqrt{n}}.$$

No caso de populações finitas com as amostras retiradas *sem reposição*, a estimativa do erro padrão da média deve ser corrigida através do fator de correção para populações finitas. Daí resulta que

$$S(\bar{X}) = \frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Podemos observar que o fator de correção ocorre no mesmo contexto que corrige a variância da distribuição binomial para a obtenção da variância da distribuição hipergeométrica.

Discutiremos a seguir a metodologia de dimensionamento de amostras para os dois casos: populações infinitas e populações finitas.

4.6.1. Dimensionamento de amostras para estimar a média de uma população

◆ **Populações infinitas (ou amostragem com reposição)**

No caso de população infinita, o intervalo de confiança para a média μ é obtido através de

$$IC(\mu;1-\alpha):\overline{x}_1 \pm t_{\alpha/2}\sqrt{\frac{s_1}{n}},$$

onde:

$\overline{x}_1$: média da pré-amostra;

s_1: desvio padrão da pré-amostra;

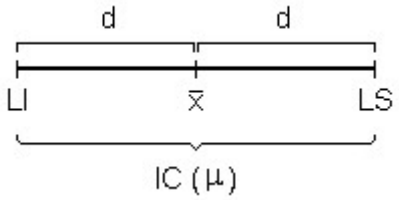
$t_{\alpha/2}$: valor crítico ao nível de confiança $1-\alpha$, correspondente ao número de graus de liberdade ($\nu_1 = n_1 - 1$) da pré-amostra;

n: tamanho suficiente da amostra.

A semi-amplitude do intervalo de confiança, portanto, é dada por $d = t_{\alpha/2}\sqrt{\frac{s_1}{n}}$. Assim, temos que

$$IC(\mu;1-\alpha):\overline{x}_1 \pm d$$

O grau de precisão é utilizado neste caso para estabelecer a semi-amplitude desejada para o intervalo de confiança. Tomando $d = \gamma\,\overline{x}_1$, com γ entre 0 e 1, estabelecemos o grau de precisão como uma percentagem da média. Por exemplo, tomar $\gamma = 0{,}10$ significa que pretendemos obter um tamanho de amostra n tal que, no intervalo de confiança para μ, tenhamos uma semi-amplitude máxima que corresponda a 10% do valor da média amostral.



Da expressão da semi-amplitude do intervalo $d = t_{\alpha/2}\sqrt{\frac{s_1}{n}}$, isolando n temos

$$\sqrt{n} = t_{\alpha/2}\frac{s_1}{d},$$

donde resulta

$$n = t_{\alpha/2}^2\frac{s_1^2}{d^2},$$

onde: n é o tamanho suficiente da amostra, se a população for infinita.

◆ **Populações finitas e amostragem sem reposição**

No caso em que a população for finita e a amostragem for feita sem reposição, o intervalo de confiança para a média μ é obtido através de

$$IC(\mu;1-\alpha): \bar{x}_1 \pm t_{\alpha/2}\sqrt{\frac{s_1}{n}}\sqrt{\frac{N-n}{N-1}},$$

onde:

$\frac{N-n}{N-1}$: é o fator de correção para populações finitas.

Nesse caso, a semi-amplitude do intervalo de confiança será $d = t_{\alpha/2}\sqrt{\frac{s_1}{n}}\sqrt{\frac{N-n}{N-1}}$. Isolando n, temos

$$\sqrt{n} = t_{\alpha/2}\frac{s_1}{d}\sqrt{\frac{N-n}{N-1}},$$

donde resulta

$$n = t^2_{\alpha/2}\frac{s_1^2}{d^2}\frac{N-n}{N-1}.$$

Fazendo $t^2_{\alpha/2}\frac{s_1^2}{d^2} = n_0$, temos

$$n = n_0\frac{N-n}{N-1}$$
$$n(N-1) = n_0N - n_0n$$
$$n(N-1) + n_0n = n_0N$$
$$n(N-1+n_0) = n_0N$$
$$n = \frac{n_0N}{N-1+n_0} = \frac{\frac{n_0N}{N}}{\frac{N-1+n_0}{N}} = \frac{n_0}{\frac{N}{N}+\frac{n_0-1}{N}} = \frac{n_0}{1+\frac{n_0-1}{N}}.$$

Como podemos observar, o tamanho ideal da amostra para populações finitas é obtido por meio de um ajuste no que seria o tamanho da amostra caso a população fosse infinita (n_0), ou seja:

$$n = \frac{n_0}{1+\frac{n_0-1}{N}}, \text{ com } n_0 = t^2_{\alpha/2}\frac{s_1^2}{d^2}.$$

A razão $\frac{n_0}{N}$ é definida como sendo a fração de amostragem. Essa quantidade pode ser utilizada para que tenhamos uma ideia da necessidade da correção para populações finitas. A regra prática é a seguinte: quando a fração de amostragem for menor que 0,05, ou seja, $\frac{n_0}{N} < 0{,}05$, significa que o tamanho da amostra representa menos de 5% do tamanho da população e, portanto, podemos considerar, para todos os efeitos, a população como sendo *infinita*. Assim, mesmo que a amostragem tenha sido feita *sem* reposição, como o tamanho da população é muito grande em relação ao tamanho da amostra, a probabilidade associada a cada elemento da população não se altera significativamente ao longo do processo de retirada. Nestes casos, não haverá a necessidade de proceder à correção para populações finitas. Observe que o princípio da correção é a diminuição do tamanho da amostra, se a população for finita. A fração de amostragem fornece uma ideia sobre a necessidade prática dessa correção.

Vejamos agora um exemplo resolvido.

O fornecimento de leite, em litros, em uma Cooperativa de Pelotas, no mês de dezembro de 1979, referente a 120 pequenos produtores, foi o seguinte:

i	Produção	i	Produção	i	Produção	i	Produção	i	Produção	i	Produção
01	150	21	265	41	285	61	285	81	450	101	380
02	140	22	140	42	380	62	290	82	354	102	400
03	400	23	320	43	170	63	500	83	420	103	140
04	250	24	290	44	164	64	500	84	140	104	290
05	200	25	100	45	380	65	350	85	470	105	264
06	400	26	300	46	420	66	500	86	280	106	175
07	400	27	450	47	194	67	140	87	100	107	285
08	400	28	150	48	444	68	194	88	450	108	260
09	450	29	240	49	300	69	300	89	230	109	430
10	285	30	194	50	240	70	300	90	450	110	270
11	500	31	274	51	470	71	400	91	300	111	374
12	280	32	400	52	100	72	250	92	400	112	400
13	150	33	200	53	380	73	500	93	290	113	192
14	350	34	350	54	474	74	300	94	284	114	310
15	194	35	100	55	400	75	230	95	300	115	450
16	450	36	500	56	343	76	194	96	420	116	480
17	200	37	120	57	350	77	330	97	150	117	300
18	474	38	234	58	404	78	250	98	435	118	344
19	170	39	400	59	474	79	284	99	270	119	284
20	284	40	400	60	340	80	434	100	400	120	260

a) Amostragem sem reposição

Utilize uma amostra preliminar de tamanho n_1= 20 e dimensione uma amostra, com grau de precisão γ = 0,10, para construir um intervalo de confiança, ao nível de 95%, para a verdadeira produção média de leite.

b) Amostragem com reposição

Utilize uma amostra preliminar de tamanho n_1 = 12 e dimensione uma amostra, com grau de precisão γ = 0,10, para construir um intervalo de confiança, ao nível de 95%, para a verdadeira produção média de leite.

Resolução:

a) Amostragem sem reposição: o processo de sorteio pode ser realizado de várias maneiras. Uma das mais simples e direta é a utilização de um gerador de números aleatórios para obtenção de quais produtores pertencerão à amostra. Como a amostragem é sem reposição, um produtor não poderá ser sorteado mais de uma vez.

Considerando a seguinte composição para a amostra preliminar sorteada de tamanho n_1 = 20:

400, 285, 280, 450, 285, 380, 290, 150, 400, 400, 170,
444, 470, 400, 200, 285, 500, 300, 284, 238,

temos:

$\bar{x}_1 = 330{,}55$ litros

$s_1^2 = 10.239{,}21$ litros^2

$\nu_1 = 19$

$\alpha = 0{,}05$

$t_{\alpha/2(19)} = 2{,}093$

$$d = \gamma \bar{x}_1 = 0{,}10 \times 330{,}55 = 33{,}06$$

$$n_0 = t^2_{\alpha/2} \frac{s_1^2}{d^2} = 2{,}093^2 \times \frac{10239{,}21}{33{,}06^2} = 41{,}05$$

Como a amostragem foi feita sem reposição e o tamanho da amostra (n) representa mais de 5% do tamanho da população (N), ou seja, $\frac{n_0}{N} = \frac{41{,}05}{120} = 0{,}3421 > 0{,}05$, a população é considerada finita. Neste caso, devemos proceder à correção de n:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} = \frac{41{,}05}{1 + \frac{41{,}05 - 1}{120}} = 30{,}77 \cong 31$$

Verificamos, assim, que o tamanho de amostra suficiente para estimar a média desta população de tamanho 120 não deve ser inferior a 31. Porém, não há necessidade de escolher uma nova amostra de tamanho 31, basta continuar com o mesmo processo e sortear mais 11 elementos que serão reunidos com os 20 da amostra preliminar.

Novo sorteio: 140, 290, 300, 450, 285, 140, 270, 374, 474, 192, 350

Amostra de tamanho suficiente: n = 31

400, 285, 280, 450, 285, 380, 290, 150, 400, 400, 170, 444, 470, 400, 200, 285, 500, 300, 284, 238, 140, 290, 300, 450, 285, 140, 270, 374, 474, 192, 350

A partir da amostra de tamanho suficiente, podemos obter o intervalo de confiança para a média da população. Os valores das estatísticas e outras quantidades de interesse são:

$\bar{x}$ = 318,58 litros, s = 104,46 litros, $\nu = 30$, $\alpha = 0{,}05$, $t_{\alpha/2(30)} = 2{,}042$

$$IC(\theta; 1-\alpha): \hat{\theta} \pm t_{\alpha/2}\, S(\hat{\theta})$$

Sendo $\theta = \mu$, temos $\hat{\theta} = \bar{X} = 318{,}58$

Como a população é finita, devemos proceder à correção do erro padrão do estimador, resultando assim

$$S(\hat{\theta}) = S(\bar{X}) = \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{104{,}46}{\sqrt{31}} \sqrt{\frac{120-31}{120-1}} = 18{,}76 \times 0{,}8648 = 16{,}22$$

IC (μ; 0,95): $318{,}58 \pm 2{,}042 \times 16{,}22$
IC (μ; 0,95): $318{,}58 \pm 33{,}13$
Limite inferior = 318,58 - 33,13 = 285,45
Limite superior = 318,58 + 33,13 = 351,71

P(285,45 < μ < 351,71) = 0,95

Concluímos que a probabilidade de o intervalo 285,45 e 351,71 litros cobrir a verdadeira produção média de leite é de 0,95.

b) Amostragem com reposição:

Amostra preliminar: n_1 =12

450, 285, 380, 290, 150, 470, 400, 200, 285, 500, 300, 284

$\bar{x}$ = 332,83 litros, s_1^2 = 11.651,79 litros, $\nu = 11$, $\alpha = 0{,}05$, $t_{\alpha/2\,(11)} = 2{,}201$

$$d = \gamma \bar{x}_1 = 0{,}10 \times 332{,}83 = 33{,}28$$

$$n = t^2_{\alpha/2} \frac{s_1^2}{d^2} = 2{,}201^2 \times \frac{11651{,}79}{33{,}28^2} = 50{,}96 \cong 51$$

Como a amostragem foi feita com reposição a população é considerada *infinita*, não havendo necessidade de proceder à correção de n.

Verificamos, assim, que o tamanho de amostra suficiente para estimar a média desta população de tamanho 120 não deve ser inferior a 51. Sorteiam-se mais 39 elementos que serão reunidos com os 12 da amostra preliminar.

Novo sorteio: 400, 285, 280, 450, 285, 380, 290, 150, 400, 400, 170, 444, 470, 400, 200, 285, 500, 300, 284, 238, 140, 290, 300, 450, 285, 140, 270, 374, 474, 192, 350, 380, 290, 150, 400, 400, 170, 444, 470

Amostra de tamanho suficiente: n = 51

450, 285, 380, 290, 150, 470, 400, 200, 285, 500, 300, 284, 400, 285, 280, 450, 285, 380, 290, 150, 400, 400, 170, 444, 470, 400, 200, 285, 500, 300, 284, 238, 140, 290, 300, 450, 285, 140, 270, 374, 474, 192, 350, 380, 290, 150, 400, 400, 170, 444, 470

A partir da amostra de tamanho suficiente, obtemos o intervalo de confiança para a média da população.

$\bar{x}$ = 324,98 litros, s = 106,10 litros, ν = 50, $t_{\alpha/2(50)} \cong t_{\alpha/2(40)}$ = 2,021

$$IC(\theta; 1-\alpha): \hat{\theta} \pm t_{\alpha/2} S(\hat{\theta})$$

Sendo $\theta = \mu$, temos:

$$\hat{\theta} = \bar{X} = 324{,}98$$

$$S(\hat{\theta}) = S(\bar{X}) = \frac{S}{\sqrt{n}} = \frac{106{,}10}{\sqrt{51}} = 14{,}86$$

IC (μ; 0,95): $324{,}98 \pm 2{,}021 \times 14{,}86$
IC (μ; 0,95): $324{,}98 \pm 30{,}03$
Limite inferior = 324,98 - 30,03= 294,95
Limite superior = 324,98 + 30,03 = 355,01

P(294,95 < μ < 355,01) = 0,95

Concluímos que a probabilidade de o intervalo de 294,95 a 355,01 litros conter a verdadeira produção média de leite é de 0,95.

4.6.2. Dimensionamento de amostras para estimar a proporção de uma população

Vimos que o intervalo de confiança para estimar a proporção populacional é obtido através da seguinte expressão:

$$IC(\pi;1-\alpha): p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

onde:

p é a estimativa da proporção populacional π;
n é o tamanho da amostra;
$z_{\alpha/2}$ é o valor da variável Z que delimita a área $\alpha/2$.

Da mesma forma que no intervalo de confiança para média, vimos que a semi-amplitude do intervalo é dada por

$$d = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}},$$

também denominado grau de precisão do intervalo. Isolando n na expressão encontramos o seguinte resultado

$$n = \frac{z^2_{\alpha/2}\, p(1-p)}{d^2},$$

onde: p é a proporção obtida a partir de uma amostra piloto.

Entretanto, é possível dimensionar o tamanho da amostra para obter a proporção p, sem fazer uma amostra piloto. Isso é feito através da superestimação de n. Quando o valor p for igual a 0,5, obtém-se o maior valor para o produto p(1−p). Assim, fazendo p = 0,5 obtemos o maior valor possível para n, garantindo a confiança e a precisão desejadas.

Exemplos resolvidos:

Exemplo 1. O fornecedor alega que entrega 10% de produtos defeituosos. Qual é o tamanho de amostra suficiente para estimar a proporção de produtos defeituosos entregues por este fornecedor, com precisão de 0,03 e 95% de confiança?

Resolução:

Temos: $z_{0,025} = 1,96$, d = 0,03 e p = 0,10.

Assim o tamanho mínimo de amostra é calculado:

$$n = \frac{z^2_{\alpha/2}\, p(1-p)}{d^2} = \frac{1,96^2 \times 0,10 \times (1-0,10)}{0,03^2} = 384,16$$

Logo, é necessária uma amostra de no mínimo 385 produtos para estimar a proporção de produtos defeituosos.

Exemplo 2. Quer-se estimar a proporção de pelotenses maiores de 16 anos que são favoráveis à flexibilização das leis trabalhistas. Qual o tamanho mínimo da amostra necessário para um erro absoluto máximo de estimação de 0,02, com um nível de confiança de 95%?

Resolução:

Temos: $z_{0,025} = 1,96$, $d = 0,02$.

Como não temos um valor prévio para p, utilizamos $p = 0,5$ para calcular o tamanho mínimo de amostra:

$$n = \frac{z_{\alpha/2}^2 p(1-p)}{d^2} = \frac{1,96^2 \times 0,05 \times (1-0,05)}{0,02^2} = 2401.$$

Portanto, é necessária uma amostra de 2401 pessoas.