4.8.2. Estimação de média (amostras grandes)

Na seção anterior, vimos como calcular o *intervalo de confiança* que nos permite fazer uma estimativa da proporção de sucessos numa população, a partir da proporção de sucessos encontrada numa amostra retirada desta população. Iremos agora, usando um raciocínio similar, ver como calcular o intervalo de confiança que irá nos permitir estimar a média de uma população, a partir da média de uma amostra.

4.8.2.1. Teorema do limite central

Na seção 4.5.3.1 vimos o *Teorema do Limite Central*, que afirma que, para amostras aleatórias grandes retiradas de uma população (de qualquer distribuição), a distribuição das médias destas amostras tende para uma distribuição normal. Reproduzimos abaixo este teorema.

Teorema do Limite Central
Se amostras aleatórias simples de tamanho *n* são retiradas de uma população (de qualquer distribuição) de média μ e desvio-padrão σ, a média amostral $\overline{X}$ será uma variável que tende para uma distribuição normal:

$$\overline{X} \rightarrow N(\mu_{\overline{X}}, \sigma^2_{\overline{X}}) \quad \text{quando } n \rightarrow \infty$$

$$\text{na qual } \mu_{\overline{X}} = \mu \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Portanto, se conhecemos os parâmetros da população (média e desvio-padrão), podemos calcular um intervalo que tenha uma probabilidade conhecida de conter a média das amostras retiradas desta população. Por outro lado, veremos agora que também é possível, se conhecemos as estatísticas de uma amostra (média e desvio-padrão), calcularmos um intervalo que tenha uma probabilidade conhecida de conter a média da população.

4.8.2.2. Construção do intervalo de confiança

O *Teorema do limite central a*firma que a média amostral $\overline{X}$ tem distribuição amostral gaussiana com parâmetros:

$$\mu_{\overline{X}} = \mu \qquad \textbf{(1)}$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \qquad \textbf{(2)}$$

Podemos então afirmar que há uma probabilidade P de a média de uma amostra aleatória qualquer se encontrar dentro do intervalo delimitado pelos valores

$$\mu_{\overline{X}} - z\sigma_{\overline{X}} \qquad \text{e} \qquad \mu_{\overline{X}} + z\sigma_{\overline{X}}$$

nos quais z é o valor da variável padronizada correspondente à probabilidade P. Se usarmos uma probabilidade $P = 0{,}95$, obtemos na tabela da curva normal o valor z = 1,96 e o intervalo se estenderá de $\mu_{\overline{X}} - 1{,}96\sigma_{\overline{X}}$ a $\mu_{\overline{X}} + 1{,}96\sigma_{\overline{X}}$, como mostrado na Fig. **1**.

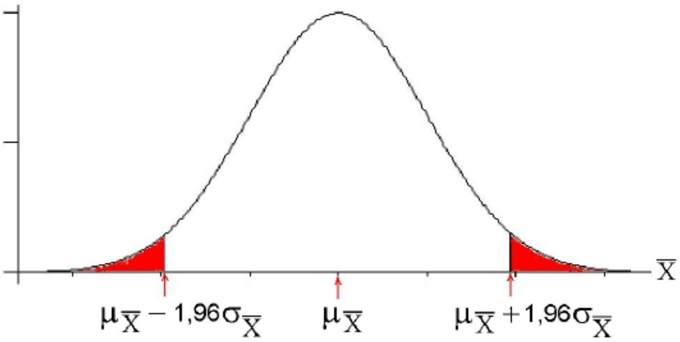

Figura 1. Intervalo com probabilidade 0,95 de conter $\overline{X}$

Podemos portanto escrever:

$$P(\mu_{\overline{X}} - 1{,}96\sigma_{\overline{X}} < \overline{X} < \mu_{\overline{X}} + 1{,}96\sigma_{\overline{X}}) \ = 0{,}95$$

Substituindo os valores dos parâmetros $\mu_{\overline{X}}$ e $\sigma_{\overline{X}}$ da distribuição amostral pelos valores em (**1**) e (**2**), dados pelo teorema do limite central, obtemos:

$$P\left[\mu - 1{,}96\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 1{,}96\frac{\sigma}{\sqrt{n}}\right] = 0.95 \qquad (\mathbf{3})$$

Se considerarmos apenas o lado esquerdo da desigualdade em (**3**):

$$\mu - 1{,}96\frac{\sigma}{\sqrt{n}} < \overline{X}$$

$$\mu < \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}} \qquad (\mathbf{4})$$

Considerando agora o lado direito:

$$\overline{X} < \mu + 1{,}96\frac{\sigma}{\sqrt{n}}$$

$$\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu \qquad (\mathbf{5})$$

Reunindo as duas desigualdade (**4**) e (**5**),

$$\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}$$

Este intervalo tem 0,95 de probabilidade de conter o valor real da média populacional μ (que é o que queremos estimar):

$$P\left[\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

Este intervalo é chamado de *Intervalo de Confiança* de 0,95 para μ. Iremos representá-lo na forma:

$$IC_{\mu}^{0.95}:\ \overline{X} \pm 1{,}96\frac{\sigma}{\sqrt{n}} \qquad (\mathbf{6})$$

4.8.2.3. Estimação do desvio-padrão da população

O intervalo definido em (**6**) não tem grande aplicação prática, porque para calculá-lo precisamos conhecer o valor do desvio-padrão populacional σ. Em geral, se não conhecemos a média de uma população, também não conhecemos seus outros parâmetros.

Nas aplicações práticas, há duas maneiras de contornar esta dificuldade. A primeira é usar uma estimativa de σ que tenha sido obtida anteriormente, por meio de outra amostra. Isto acontece, por exemplo, em problemas de controle de qualidade, onde as características de um processo de produção (por exemplo, as dimensões de uma peça) são medidas várias vezes por dia, em intervalos regulares, e são feitas constantemente novas estimativas dos vários parâmetros que interessam.

A segunda maneira, usada mais freqüentemente, é usar o desvio-padrão da amostra (s) como uma estimativa do desvio-padrão da população (σ). Isto equivale a fazer uma estimativa *pontual* de σ, para permitir uma estimativa *intervalar* de μ. Fazemos então:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

O intervalo de confiança passa a ser escrito na forma:

$$IC_{\mu}^{0,95}:\ \overline{X} \pm 1{,}96\frac{s}{\sqrt{n}} \qquad (\mathbf{7})$$

Este processo de fazer uma estimativa dentro de outra estimativa não introduz um erro considerável, se as amostras forem grandes. Se as amostras forem pequenas, porém, o desvio-padrão s deixa de ser um bom estimador de σ, e o Teorema do Limite Central perde a validade; vimos este problema nos testes com amostras pequenas (seção 4.7.1), e tornaremos a vê-lo mais adiante, na estimação com amostras pequenas (seção 4.8.4).

4.8.2.4. Observações sobre a estimação da média por IC

Quando discutimos a estimação da proporção de sucessos de uma população, fizemos algumas observações sobre os intervalos de confiança (seção 4.8.1.3). As mesmas observações, com alguns ajustes, são válidas para a estimação da média por ICs, como veremos a seguir.

(i) Nível de confiança do IC

O intervalo definido em (**7**) tem um *nível de confiança* de 0,95, o que significa que há uma probabilidade de 0,95 de ele conter o valor real de μ. Podemos, é claro, fazer intervalos cujo nível de confiança seja diferente de 0,95; para isto, teremos que trocar o valor de z = 1,96 pelo valor correspondente à probabilidade desejada. Além de 0,95, também são usados às vezes níveis de 0,90 ou de 0,99, aos quais correspondem os valores z = 1,645 e z = 2,575, respectivamente. Os ICs são então dados pelas expressões:

$$IC_{\mu}^{0.90}: \ \overline{X} \pm 1{,}645 \frac{s}{\sqrt{n}}$$

$$IC_{\mu}^{0.99}: \ \overline{X} \pm 2{,}575 \frac{s}{\sqrt{n}}$$

(ii) Margem de erro e determinação do tamanho da amostra

Em (**7**), o termo

$$1{,}96 \frac{s}{\sqrt{n}} \qquad (\mathbf{8})$$

dá a *margem de erro* da estimativa. O tamanho desta margem depende da variância *s* da amostra, sobre a qual não temos nenhum controle, e do tamanho *n* da amostra, que podemos controlar. Note que o *n* está no denominador da fração em (**8**), o que significa que quanto maior a amostra que usarmos, menor será a margem de erro. Contudo, como também acontece nos ICs para proporção, o *n* está dentro de um radical, e a margem de erro não é portanto uma função linear do tamanho da amostra: para reduzir à metade a margem de erro, por exemplo, teremos que usar uma amostra quatro vezes maior.

Se desejamos estimar a média de uma população por um IC com margem de erro dada, podemos usar (**8**) para calcular qual será o tamanho necessário da amostra. Por exemplo, se queremos estimar uma das dimensões das peças produzidas por uma fábrica por um IC com margem de erro igual a ± 0,05 mm, podemos fazer

$$1{,}96 \frac{s}{\sqrt{n}} = 0{,}05 \qquad \rightarrow \qquad \left(1{,}96 \frac{s}{0{,}05}\right)^2 = n$$

Precisamos do valor do desvio-padrão *s* da amostra. Podemos usar como estimativa deste *s* o desvio-padrão calculado anteriormente em outra amostra (no controle de qualidade, amostras dos itens produzidos são retiradas freqüentemente, a intervalos regulares). Ou então, se não há estimativas anteriores, podemos simplesmente retirar uma amostra não muito grande e usar seu desvio-padrão para calcular qual deve ser, pelo menos aproximadamente, o *n* necessário para atingir a margem de erro desejada. Estas amostras relativamente pequenas que retiramos no início de uma pesquisa, para termos uma idéia das características de uma população, são chamadas de *amostras-piloto*; neste tipo de problema, serviriam por exemplo para calcularmos o tamanho *n* da amostra, e a partir daí, estimarmos qual seria o custo, em termos de tempo e de dinheiro, para obtermos a estimativa com a margem de erro desejada.

(iii) Interpretação do IC

Se calculamos um IC com um nível de confiança de 0,95 usando a expressão (**7**), há uma probabilidade de 0,95 de ele conter o valor real da média μ da população. Isto quer dizer que, se tirarmos um grande número de amostras de uma população e calcularmos estes ICs a partir delas, podemos esperar que em média 95% deles (19 em cada 20 ICs) irão conter o valor real de μ; os outros 5%, porém (1 em cada 20 ICs) darão estimativas erradas, pois os intervalos não conterão o valor real de μ.

Para fazer o gráfico da Fig. **2**, retiramos 50 amostras aleatórias de tamanho n = 50, de uma população simulada com distribuição normal de média μ=5 e variância σ^2=1, e calculamos ICs a partir de cada uma delas. O valor real de μ está destacado na linha horizontal vermelha. Os ICs têm amplitude diferentes entre si (indicada pela altura dos retângulos), porque as margens de erros de cada IC foram calculadas a partir dos desvios-padrões *s* de cada amostra. Note que destes 50 ICs, 47 (94%) contém o valor verdadeiro da média (μ=5); há porém três ICs (6%), destacados em azul, que não contém este valor e são portanto estimativas errôneas.

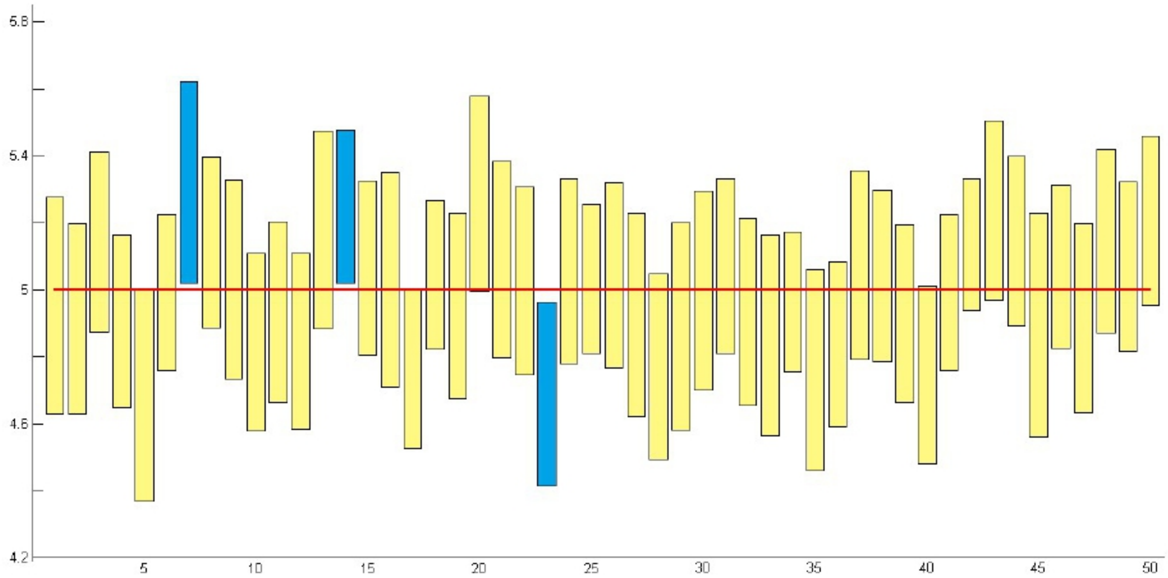

Figura 2. Intervalos de confiança calculados a partir de amostras simuladas de uma população N(5,1)

Quando retiramos uma amostra de uma população e estimamos a partir dela a média μ por meio de um IC, nunca poderemos saber se a amostra que usamos foi uma das que produzem estimativas corretas, ou uma das que produzem estimativas errôneas. É mais provável seja daquelas que dão estimativas corretas (probabilidade = 0,95), mas nunca poderemos ter certeza. Como em qualquer técnica de Inferência Estatística, também na estimação sempre existe a probabilidade de chegarmos a um resultado errado; podemos reduzir esta probabilidade, mas nunca eliminá-la.

(iv) É preciso levar em conta o tipo de amostra

É preciso ter sempre em mente um ponto importante: todas estas fórmulas acima, derivadas do Teorema do Limite Central *só servem se a amostra for grande e aleatória simples.* Se a amostra pequena, ou foi retirada por meio de alguma outra técnica de amostragem (amostras *estratificadas*, por *conglomerados*, etc.; veja seção 5.1), a distribuição amostral da média amostral $\overline{X}$ será diferente da estipulada pelo teorema, e teremos que usar outras fórmulas.

4.8.2.5. Exemplo

Voltemos ao exemplo mostrado anteriormente (seções 4.5.2 e 4.5.3), cujo enunciado repetimos abaixo:

> Devido ao excesso de pesca, surgiu a hipótese de que o peso médio dos peixes de uma espécie comum no Atlântico Norte está diminuindo, porque estes peixes não têm tempo de crescer suficientemente antes de serem pescados. Esta espécie tinha anteriormente peso médio de 28,0 kg, com desvio padrão de 4,0 kg. Um pesquisador retira uma amostra aleatória de 60 destes peixes, e encontra um peso médio de 26,0 kg. Este resultado é uma evidência de que o peso dos peixes está diminuindo?

Na seção 4.5.3.2, fizemos um teste de hipótese a partir desta amostra e concluímos que realmente existe evidência de que o peso médio dos peixes não é mais μ=28 kg (p = 0,00005), e sim algum valor menor que 28 kg. Iremos agora estimar este valor, por meio de um IC de 0,95 de confiança.

O tamanho da amostra e as estatísticas amostrais fornecidas são:

$$n = 60$$
$$s = 4{,}0 \text{ kg}$$
$$\overline{X} = 26{,}0 \text{ kg}$$

Usando a expressão (7),

$$IC_{\mu}^{0,95}: \ \overline{X} \pm 1{,}96\frac{s}{\sqrt{n}}$$

$$IC_{\mu}^{0,95}: \ 26{,}0 \pm 1{,}96\frac{4{,}0}{\sqrt{60}}$$

$$IC_{\mu}^{0,95}: \ (24{,}99 \ \ a \ \ 27{,}01)\,kg$$

O IC calculado não contém o valor hipotético da média, μ = 28 kg; este IC, portanto, corrobora a conclusão que tínhamos obtido por meio do teste de hipóteses: a média atual do peso é realmente *menor* que 28,0 kg.