4.8.4. Estimação de médias (amostras pequenas)

A estimação da *média* de uma população por meio de intervalos de confiança com amostras pequenas segue um procedimento similar ao que já vimos para a estimação com amostras grandes (seção 4.8.2); a diferença é que os cálculos agora não serão baseados no modelo de distribuição *normal*, e sim no modelo de distribuição de *Student*.

4.8.4.1. Intervalo de confiança para a média

Vimos anteriormente um teorema afirmando que, se uma população pode ser considerada normal e tem desvio-padrão σ conhecido, a média $\overline{X}$ das amostras aleatórias simples dela retiradas será uma variável com distribuição normal, qualquer que seja o tamanho da amostra (Teorema 1, seção 4.7.1.1). Este teorema está reproduzido abaixo.

Teorema 1. *Distribuição das médias $\overline{X}$ de amostras de população normal*
Se amostras aleatórias simples de tamanho *n* (qualquer) são retiradas de uma população normal de média μ e desvio-padrão σ, a média amostral $\overline{X}$ será uma variável com distribuição normal,

$$\overline{X} \sim N(\mu_{\overline{X}}, \sigma^2_{\overline{X}})$$

cujos parâmetros são: $\mu_{\overline{X}} = \mu$ e $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$

A variável padronizada $Z \sim N(0,1)$ será dada por

$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \rightarrow z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \qquad (1)$$

Podemos portanto fazer testes de hipóteses, ou calcular intervalos de confiança, usando a distribuição normal. Vimos porém outro teorema afirmando que, se esta mesma população tiver desvio-padrão desconhecido, a média $\overline{X}$ das amostras aleatórias simples dela retiradas não será uma variável com distribuição normal, mas poderá ser transformada numa variável *t*, que segue a distribuição de *Student*. Este teorema (Teorema 2, seção 4.7.1.1) está reproduzido abaixo:

Teorema 2. Estatística de teste *t*, com distribuição de Student
Se amostras aleatórias simples de tamanho *n* são retiradas de uma população de distribuição normal de média μ, a estatística de teste *t* definida como

$$t = \frac{\overline{X} - \mu}{s_{corr} / \sqrt{n}} \qquad (2)$$

onde $\overline{X}$: média da amostra, e s_{corr} : desvio-padrão corrigido,
terá distribuição de *Student*, com $g.l. = n - 1$.

Este teorema foi a base teórica para os *testes* de médias e de diferenças de médias (seção 4.7.1), e será também a base para a *estimação* das médias e diferenças de médias.

Para amostras grandes, a fórmula para o IC de uma média é:

$$IC_{\mu}^{1-\alpha}: \ \overline{X} \pm z\frac{s}{\sqrt{n}} \qquad (3)$$

Para um IC de 0,95 (que foi o que usamos em todos os exemplos), $z = 1{,}96$, e a fórmula se torna (seção 4.8.2.3, eq. 7):

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

Quando a amostra é pequena, porém, o desvio-padrão s deve ser substituído por sua versão corrigida s_{corr}:

$$s_{corr} = \sqrt{\frac{\sum(x-\overline{X})^2}{n-1}} \qquad (4)$$

e a variável z tem que ser substituída por t. O valor de t depende do número de graus de liberdade existentes (definido pelo tamanho n da amostra) e do *nível de confiança* do intervalo (representado por $1\text{-}\alpha$; lembre-se de que α é a probabilidade *fora* do intervalo desejado, e corresponde à *região de rejeição* num teste de hipótese). Se queremos fazer um IC com $1\text{-}\alpha$ de confiança usando uma amostra pequena, o intervalo será:

$$IC_{\mu}^{1-\alpha} = \overline{X} \pm t_{(1-\alpha,g.l.)}\frac{s_{corr}}{\sqrt{n}} \qquad (5)$$

4.8.4.2. Os problemas das amostras pequenas

Nenhum dos dois teoremas acima menciona o *tamanho* da amostra, que parece não ser relevante. Por que então temos que usar o Teorema **2** quando a amostra for pequena, mas não quando ela for grande? (Note que fizemos THs e ICs para amostras grandes nas seções 4.4, 4.5, e 4.6 sem nos preocuparmos com isto, usando sempre a distribuição normal).

A razão é que, quando as amostras são grandes, tudo fica mais fácil. Primeiro, a exigência de que a população seja normal (comum aos dois teoremas acima) pode ser relaxada; se a amostra for suficientemente grande, a distribuição amostral de $\overline{X}$ *tenderá* para normal, mesmo que a população não seja normal (este é o *Teorema do Limite Central*, visto na seção 4.5.3.1, Teorema 2). Segundo, se o desvio-padrão σ da população foi desconhecido, poderá ser estimado simplesmente pelo desvio-padrão s da amostra, e ainda assim a distribuição amostral de $\overline{X}$ continuará sendo normal.

Quando as amostras são *pequenas*, porém, tudo se complica, pois estas amostras trazem menos informação do que as amostras grandes. Se a população não for normal, a distribuição amostral de $\overline{X}$ não irá se aproximar suficientemente da normal. Se a população for normal, mas o desvio-padrão σ da população for desconhecido, teremos que estimar este σ a partir do desvio-padrão s_{corr} da amostra; serão feitas portanto duas estimativas (de σ e de $\overline{X}$) ao invés de uma só, o que aumenta a incerteza do resultado. Isto faz com

que os testes com amostras pequenas tenham sempre menor *poder* do que os feitos com amostras grandes (veja "poder" nas seções 4.3.3 e 4.7.1.1, p.3); além disso, que os ICs calculados a partir destas amostras tenham maior amplitude do que os calculados a partir da distribuição normal, e sejam portanto menos *precisos*.

Por exemplo, suponha que queiramos estimar a média μ de uma população normal, e para isto retiramos dela uma amostra de $n = 10$ e encontramos $\overline{X} = 2{,}00$. Se sabemos que o desvio-padrão desta população é $\sigma = 0{,}5$, o IC de 95% de confiança pode ser calculado por meio da distribuição normal, e resulta em:

$$IC^{\mu}_{0,95} = \overline{X} \pm 1{,}96\frac{\sigma}{\sqrt{n}}$$

$$IC^{\mu}_{0,95} = 2{,}00 \pm 1{,}96\frac{0{,}5}{\sqrt{10}} = 2{,}00 \pm 0{,}30 \qquad (6)$$

Se contudo σ é desconhecido, teremos que estimá-lo a partir do desvio-padrão da amostra. Suponha que na amostra encontramos $s_{corr} = 0{,}5$. O IC calculado a partir da distribuição *t* será dado por:

$$IC^{\mu}_{0,95} = \overline{X} \pm t_{(1-\alpha=0,95;\, g.l.)}\frac{s_{corr}}{\sqrt{n}}$$

Na tabela da distribuição de Student, vemos que o valor crítico de *t* (para $g.l. = n - 1 = 9$ e probabilidade $1\text{-}\alpha = 0{,}95$) é:

$t_{(0,95;\, g.l.=9)} = 2{,}262$

$$IC^{\mu}_{0,95} = 2{,}00 \pm 2{,}262\frac{5{,}0}{\sqrt{10}} = 2{,}00 \pm 0{,}36 \qquad (7)$$

Note que, embora em ambos intervalos o valor numérico do desvio-padrão usado tenha sido o mesmo (igual a 0,5), o IC na eq. (6) tem amplitude menor do que o na eq. (7), e é portanto mais preciso. Isto é fácil de entender, intuitivamente: a estimativa na eq. (6) foi calculada a partir do valor *verdadeiro* de σ; a estimativa na eq. (7) foi calculada a partir de uma *estimativa* de σ, obtida a partir de uma amostra pequena; esta estimativa é apenas uma aproximação do valor real, e que com certeza conterá um erro que não podemos determinar.

4.8.4.3. Demonstração do intervalo de confiança

É fácil ver que a expressão do IC para amostras pequenas na eq. (5) tem a mesma estrutura do IC para amostras grandes na eq. (3), sendo feita apenas a substituição do desvio-padrão amostral *s* por sua versão corrigida s_{corr}, e da variável *z* pela variável *t*. Estas duas substituições fazem sentido, se considerarmos o que foi dito na seção anterior, e provavelmente a maioria dos estudantes não irá se preocupar com uma demonstração da origem da expressão em (5).

Esta demonstração contudo não é difícil. Primeiro, vimos na eq. (2) que a variável *t* é definida como:

$$t = \frac{\overline{X} - \mu}{s_{corr}/\sqrt{n}}$$

Representaremos os valores de t que delimitam o intervalo de confiança por t_c (t crítico); alguns livros preferem usar a notação $t_{(1-\alpha,\ g.l.)}$, mas usaremos aqui uma forma mais simples. A probabilidade do intervalo limitado pelos valores $-t_c$ e $+t_c$ será a *confiança* do intervalo, e a representaremos por $1-\alpha$:

$$P[-t_c < t < t_c] = 1-\alpha$$

Substituindo t por sua definição, e reorganizando os termos da equação, obtemos:

$$P\left[-t_c < \frac{\overline{X} - \mu}{s_{corr}/\sqrt{n}} < t_c\right] = 1-\alpha$$

$$P\left[-t_c \frac{s_{corr}}{\sqrt{n}} < \overline{X} - \mu < t_c \frac{s_{corr}}{\sqrt{n}}\right] = 1-\alpha$$

$$P\left[-\overline{X} - t_c \frac{s_{corr}}{\sqrt{n}} < -\mu < -\overline{X} + t_c \frac{s_{corr}}{\sqrt{n}}\right] = 1-\alpha$$

$$P\left[\overline{X} + t_c \frac{s_{corr}}{\sqrt{n}} > \mu > \overline{X} - t_c \frac{s_{corr}}{\sqrt{n}}\right] = 1-\alpha$$

$$P\left[\overline{X} - t_c \frac{s_{corr}}{\sqrt{n}} < \mu < \overline{X} + t_c \frac{s_{corr}}{\sqrt{n}}\right] = 1-\alpha \qquad (8)$$

A expressão em (8) dá o IC para a média populacional : um intervalo que tem uma probabilidade $1-\alpha$ de conter o valor de μ. Uma observação interessante é a de que, na maioria das expressões deste tipo que já usamos, a variável encontra-se no centro destas desigualdades; por exemplo, quando escrevemos, num problema simples de probabilidades:

$$P(3 < X < 5)$$

Em (8), porém, quem está no centro é μ, uma constante; o que é variável são os limites do intervalo,

$$\overline{X} \pm t_c \frac{s_{corr}}{\sqrt{n}}$$

pois cada amostra que tirarmos terá uma média $\overline{X}$ e um desvio-padrão s_{corr} diferente.

4.8.4.4. Verificação dos pressupostos

Os ICs para médias feitos a partir de amostras pequenas, como vimos, são bastante semelhantes aos ICs feitos a partir de amostras grandes (seção 4.8.2). É importante porém

lembrar que, quando trabalhamos com amostras pequenas, tanto em testes quanto em ICs, há sempre vários pressupostos que devem ser levados em conta; antes de fazer o IC, é preciso verificar se estes pressupostos foram atendidos.

Estes pressupostos já foram discutidos quando estudamos os testes com amostras pequenas (seção 4.7.1). Em resumo, são dois os mais importantes. O primeiro é o de que a amostra usada seja uma amostra *aleatória simples*. Se a amostragem foi feita por outra técnica (amostragem *estratificada*, amostragem *por conglomerados*, etc.), as fórmulas para cálculo da variância serão diferentes. Este pressuposto não pode ser verificado a partir da amostra. Não é possível descobrir, simplesmente analisando os dados, que tipo de amostragem foi usada; é preciso verificar como foi feito o planejamento do estudo, como os dados foram obtidos, etc.

O segundo pressuposto é o de que a distribuição da população seja normal. Isto pode ser verificado a partir da amostra, por meio de técnicas que comparam graficamente a distribuição observada nos dados com a distribuição normal teórica (*gráficos de separatrizes*), ou por meio de testes de *normalidade*, testes de hipótese que verificam se é provável que uma amostra tenha saído de uma população normal. Os testes mais usados são os de *Kolmogorov-Smirnov* e de *Shapiro,* ambos implementados em R no pacote *stats* (seção 4.7.1.2 (i)).

4.8.4.5. Exemplo

Um grupo de pesquisadoras coletou num lago uma amostra de 15 girinos de uma espécie de sapo. Os dados abaixo mostram os pesos destes girinos, em gramas:

```
2.3 1.8 2.0 2.1 2.2 2.3 2.1 2.0 2.1 2.4 2.1 2.3 2.1 2.8 2.5
```

As estatísticas amostrais calculadas são:

média: $\overline{X} = 2{,}207$ g

desvio-padrão: $s_{corr} = 0{,}240$ g

Queremos estimar o peso médio destes girinos por meio de um IC de 0,95 de confiança. Como a amostra é pequena, faremos um teste baseado na distribuição de Student. Para isto, precisamos primeiramente verificar se a população de onde saiu esta amostra é normal; o teste de Shapiro feito no R dá como resultado:

```
> shapiro.test(x)
Shapiro-Wilk normality test
data:  x
W = 0.9281, p-value = 0.2552
```

Como o valor-p obtido foi bem maior do que 0,05, concluímos que a hipótese nula (de que a população é normal) não pode ser rejeitada, e podemos prosseguir com os cálculos.

Para um IC de 0,95, precisamos do valor de t definido para:

$1-\alpha = 0{,}95$

$g.l. = n - 1 = 15 - 1 = 14$

Na tabela de Student, encontramos $t = 2{,}14$. Fazendo os cálculos,

$$IC_{\mu}^{0,95} = \overline{X} \pm t_{(0,95,g.l.=14)} \frac{s_{corr}}{\sqrt{n}}$$

$$IC_{\mu}^{0,95} = 2{,}207 \pm 2{,}14 \times \frac{0{,}240}{\sqrt{15}} = 2{,}207 \pm 0{,}133$$

O intervalo de confiança desejado é portanto:

$$IC_{\mu}^{0,95}: \text{ (2,07 a 2,34) g}$$

Resumo

1. Estimativas de médias populacionais por meio de intervalos de confiança podem ser feitas a partir de amostras pequenas; será usada então a distribuição de *Student*, em vez da distribuição normal.
2. Para calcular o IC precisamos de uma estimativa do desvio-padrão da população; esta estimativa será dada pelo *desvio-padrão corrigido* calculado na amostra.
3. Todo IC (e todo TH) feito a partir de amostras pequenas é baseado em alguns pressupostos; é preciso verificar se eles foram atendidos, antes de fazer os cálculos.
4. O IC para uma média populacional, além de pressupor que a distribuição na população seja *normal*, exige também que a amostra usada seja *aleatória simples*. A normalidade pode ser verificada a partir dos por meio de gráficos ou de *testes de normalidade*. Contudo, não é possível verificar a partir dos dados se a amostra foi aleatória simples.
5. Assim como todo TH feito com amostra pequena tem menos *poder* do que os THs feitos com amostras grandes, todo IC feito com amostras pequenas será sempre menos *preciso* do que os ICs feitos com amostras grandes.