Resumo 10 – Estimação de parâmetros populacionais

9.1. Introdução

Aqui estudaremos o problema de avaliar certas características dos elementos da população (parâmetros), com base em operações com os dados de uma amostra (estatística).

Na estimação de parâmetros fazemos um raciocínio tipicamente indutivo, porque generalizamos resultados da amostra para a população.

Reforçando alguns conceitos:

População: No sentido geral, é um conjunto de elementos com pelo menos uma característica em comum. Essa característica deve delimitar, inequivocamente, quais elementos pertencem a população e quais não pertencem.

Parâmetro: é uma medida que descreve certa característica numérica dos elementos da população.

Amostra aleatória simples: uma parte da população, sendo que os elementos são extraídos por sorteio.

Estimador ou Estatística de um parâmetro: alguma medida associada com os dados de uma amostra a ser extraída da população. Quando usada com o objetivo de avaliar (estimar) o valor de algum parâmetro, também é chamada de estimador do parâmetro.

Estimativa: é o valor da estatística (estimador), calculado com base na amostra efetivamente observada.

Erro amostral: é a diferença entre uma estatística e o parâmetro que se quer estimar.

Exemplo 9.1: A prefeitura pretende avaliar a aceitação de um projeto de mudança no transporte coletivo. Depois de apresentá-lo aos usuários, os responsáveis por sua execução pretendem conhecer, mesmo que de forma aproximada, o parâmetro p (proporção de favoráveis ao projeto na população de usuários do transporte coletivo do município.

Para estimar este parâmetro, a prefeitura planeja uma amostragem aleatória simples de 400 usuários. Dessa amostra, calcular a estatística $\hat{p}$ (proporção de moradores favoráveis ao projeto na amostra.

Observada efetivamente a amostra, devemos ter $p \neq \hat{p}$, devido ao erro amostral.

Exemplo 9.2: Para estudar o efeito da merenda escolar, introduzida nas escolas do município, planeja-se acompanhar um amostra de 100 crianças, que estão entrando na rede municipal de ensino. Dentre diversas características de interesse, pretende-se avaliar o parâmetro μ (ganho médio de peso durante o primeiro ano letivo na população de crianças da rede municipal de ensino.

Da amostra de crianças em estudo, pode-se calcular a estatística $\bar{X}$ (ganho médio de peso, durante o primeiro ano letivo, das 100 crianças em observação.

A estatística $\bar{X}$ pode ser usada como um estimador do parâmetro μ mas, como no exemplo anterior, devemos ter $\mu \neq \bar{X}$ devido ao erro amostral.

Nas próximas seções estudaremos um processo que permite avaliar a margem de erro que podemos estar cometendo por examinar apenas uma amostra e não a população toda.

Quando estivermos estudando a incidência de algum atributo numa certa população, geralmente o interesse esta na proporção ou porcentagem de elementos com o atributo, como no Exemplo 9.1. Quando estamos pesquisando alguma característica quantitativa, como no Exemplo 9.2, é comum o interesse em estimar uma média ou um desvio padrão.

A seguir, temos alguns parâmetros e as respectivas estatísticas que geralmente são usadas para estimá-los.

Parâmetros (característica da população)	Estatística (característica da amostra)
p = proporção de algum atributo, dentre os elementos da população	$\hat{p}$ = proporção de elementos com o atributo dentre os que serão observados na amostra.
μ = média de alguma variável, dentre os elementos da população	$\bar{X}$ = média da variável, a ser calculada com os elementos da amostra.
σ = desvio padrão de uma variável. Dentre os elementos da população	S = desvio padrão da variável a ser calculado com os elementos da amostra.

Em geral, parâmetros são desconhecidos. As **estatísticas** são **variáveis aleatórias**, pois seus valores dependem dos elementos a serem sorteados na amostragem. Ao se observar uma amostra a **estatística** se identifica com um valor (resultado do cálculo), chamado de **estimativa**.

Um dos principais objetivos na teoria da estimação é estimar um **limite superior provável** para o erro amostral. Esse valor será a base para avaliarmos a **precisão** de nossa estimativa.

Importante: Quanto menor for o limite superior provável, mais precisa é a nossa estimativa.

9.2. Distribuição Amostral

Considere a seguinte pergunta referente ao Exemplo 9.1:

O valor de $\hat{p}$ vai ser um valor próximo da verdadeira proporção p, a qual se refere a todos os usuários do município???

Como na prática, o valor de p é desconhecido, tentamos responder de forma indireta, através do conhecimento de como se distribuem os possíveis valores de $\hat{p}$.

Diferentes valores de $\hat{p}$ podem ser obtidos de amostras diferentes da mesma população. Para cada amostra observada temos um valor $\hat{p}$.

Teorema 1: ***Teorema Central do Limite*** **(T.C.L.)**

Seja $X_1, X_2, ..., X_n$, uma seqüência de n variáveis aleatórias independentes e identicamente distribuídas (v.a.i.i.d.) com média $E(X_i) = \mu$ e variância $Var(X_i) = \sigma^2$. Então, $Y = \sum_{i=1}^{n} X_i$ converge em distribuição para uma normal com média, μ_Y, igual à $n\mu$ e variância, σ^2_Y, igual à $n\sigma^2$ quando $n \to \infty$, ou seja,

$$\cdot\ Z = \frac{[\sum_{i=1}^{n}(X_i) - n\mu]}{\sqrt{n\sigma^2}} \quad \begin{matrix} n\to\infty \\ \to \end{matrix} \quad N(0;1)$$

Ou ainda,

Sendo $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{Y}{n}$, temos que: $Z = \frac{Y - n\mu}{\sqrt{n\sigma^2}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \to N(0;1)$

9.2.1. *Distribuição Amostral da Proporção*

Vamos considerar uma população em que a proporção de elementos portadores de uma certa característica é *p*.

Define-se como $\hat{p}$ a proporção de indivíduos portadores da característica na amostra, isto é,

$$\hat{p} = \frac{\boldsymbol{X}}{n},$$

onde ***X*** é o número de indivíduos portadores da característica na amostra.

Pelo *Teorema Central do Limite,* $\hat{p} = \frac{\boldsymbol{X}}{n}$. terá distribuição aproximadamente normal, com média *p* e variância $\frac{p(1-p)}{n}$, ou seja,

$$\hat{p} \sim N(\mu_{\hat{p}} = p;\ \sigma^2_{\hat{p}} = \frac{p(1-p)}{n}).$$

9.2.2. Amostragem da Distribuição Normal

A distribuição normal possui dois parâmetros, μ e σ^2, e geralmente, estes são desconhecidos. Sendo $\overline{X}_n$ o melhor estimador do parâmetro μ e S^2 o melhor estimador do parâmetro σ^2 é necessário conhecer as distribuições de $\overline{X}_n$ e de S^2.

Teorema 2: Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de uma variável X, com esperança $E(X) = \mu$ e variância $Var(X) = \sigma^2$. Seja também, $\overline{X}_n$ a média amostral e S^2 a variância amostral. Então,

(i) Para amostragem com reposição de uma população finita ou qualquer tipo de amostragem de uma população infinita, temos que:

$$E(\overline{X}_n) = \mu \ , \ Var(\overline{X}_n) = \frac{\sigma^2}{n}$$

e

$$E(S^2) = \sigma^2 \ , \ Var(S^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right) \ para \ n > 1$$

onde $\sigma^4 = \left(\sigma^2\right)^2$ e $\mu_4 = E\left[\left(X-\mu\right)^4\right]$.

(ii) Para amostragem sem reposição de uma população finita, temos que:

$$Var(\overline{X}_n) = \frac{\sigma^2}{n} \times \frac{N-n}{N-1} \ ,$$

onde $\frac{N-n}{N-1}$ é chamado de *fator de correção de população finita*.

Teorema 3: ***Propriedade Aditiva da Distribuição Normal***

Seja $X_1, X_2, ..., X_n$, uma seqüência de n variáveis aleatórias independentes com distribuição $N\left(\mu_i ; \sigma_i^2\right)$, i= 1, 2, ..., n. Façamos $W = X_1 + X_2 + ... + X_n$. Então, W terá a distribuição $N\left(\sum_{i=1}^{n} \mu_i ; \sum_{i=1}^{n} \sigma_i^2\right)$.

Conseqüência: Seja $X_1, X_2, ..., X_n$, uma seqüência de n variáveis aleatórias independentes e identicamente distribuídas com distribuição $N(\mu; \sigma^2)$. Façamos $\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$. Então, $\overline{X}_n$ terá a distribuição $N\left(\mu ; \frac{\sigma^2}{n}\right)$.

Definição 1: ***Estimador*** **e** ***Estimativa***

Um *estimador* do parâmetro θ é qualquer função das observações $X_1, X_2, ...,X_n$, isto é, $g(X_1, X_2, ...,X_n)$. O valor que g assume, isto é, $g(x_1, x_2, ..., x_n)$, será referido como uma *estimativa* de θ e é usualmente escrito assim: $\hat{\theta} = g(x_1, x_2, ..., x_n)$.

Note que, segundo está definição, um estimador é qualquer estatística cujos valores são usados para estimar θ (ou uma função de θ).

O problema da estimação é, então, determinar uma função $T = g(X_1, X_2, ...,X_n)$ que seja "próxima" de θ, segundo algum critério.

Notação: θ : parâmetro a ser estimado
T : um estimador de θ
$\hat{\theta}$: uma estimativa de θ

Definição 2: ***Estimador Pontual***

Seja $X_1, X_2, ...,X_n$ uma amostra aleatória de uma variável aleatória X que descreve uma característica de interesse de uma população com uma distribuição $f_x(x;\theta)$. Então, qualquer estatística $T = g(X_1, X_2, ...,X_n)$ é um *estimador pontual* de θ.

Notação: $\hat{\theta} = T(x) = g(x_1, x_2, ..., x_n)$ é a estimativa pontual de θ.

Definição 3 - ***Intervalos de Confiança (I.C.)*****:** Seja $(X_1, X_2, ..., X_n)$ uma amostra aleatória de uma população e θ o parâmetro de interesse. Se T um estimador de θ, e conhecida distribuição amostral de T, sempre será possível achar dois valores t_1 e t_2, tal que

$$Pr(t_1 < \theta < t_2) = 1 - \alpha = \gamma$$

sendo γ um valor fixado e $0 < \gamma < 1$.

Para uma dada amostra, teremos dois valores fixos t_1 e t_2, e o intervalo de confiança para θ com nível de confiança γ será indicado do seguinte modo:

$$IC(\theta: \gamma) = [t_1, t_2]$$

9.3. Intervalos de Confiança para Parâmetros da Distribuição Normal

9.3.1. Intervalo de Confiança para μ com $\sigma^2 = \sigma_0^2$ conhecido

O intervalo de confiança para μ com 100γ % de confiança é dado por:

$$I.C.(\mu : \gamma) = \left[\overline{X} - z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}}, \ \overline{X} + z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}} \right],$$

com $Pr(Z < - z) = Pr(Z > z) = \frac{\alpha}{2}$.

9.3.2. Intervalo de Confiança para μ com σ^2 Desconhecido

O intervalo de confiança para μ com 100γ % de confiança é dado por:

$$I.C.(\mu : \gamma) = \left[\overline{X} - t \frac{S}{\sqrt{n}}, \ \overline{X} + t \frac{S}{\sqrt{n}} \right],$$

com $Pr(t_{(n-1)} < - t) = Pr(t_{(n-1)} > t) = \frac{\alpha}{2}$.

9.3.3. Intervalo de Confiança para σ^2 com μ Desconhecido

O intervalo de confiança para σ^2 com 100γ % de confiança é dado por:

$$I.C.(\sigma^2 : \gamma) = \left[\frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{q_2}, \ \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{q_1} \right] = \left[\frac{(n-1) S^2}{q_2}, \ \frac{(n-1) S^2}{q_1} \right],$$

onde q_1 e q_2 são tais que, $Pr(\chi^2_{(n-1)} < q_1) = Pr(\chi^2_{(n-1)} > q_2) = \frac{\alpha}{2}$.

9.3.4. Intervalo de Confiança para σ^2 com $\mu = \mu_0$ Conhecido

O intervalo de confiança para σ^2 com 100γ % de confiança é dado por:

$$I.C.(\sigma^2 : \gamma) = \left[\frac{\sum_{i=1}^{n} (X_i - \mu_0)^2}{q_2}, \frac{\sum_{i=1}^{n} (X_i - \mu_0)^2}{q_1} \right],$$

onde q_1 e q_2 são tais que, $Pr(\chi^2_{(n)} < q_1) = Pr((\chi^2_{(n)} > q_2) = \frac{\alpha}{2}$.

9.4. Intervalo de Confiança para Proporção Populacional

O intervalo de confiança para *p* com 100γ % de confiança é dado por:

$$I.C.(p : \gamma) = \left[\hat{p} - z_{\alpha/2} \sqrt{p(1-p)/n} ; \hat{p} + z_{\alpha/2} \sqrt{p(1-p)/n} \right]$$

Como não conhecemos *p*, usamos o fato de que $p(1-p) \leq 1/4$, logo $\sqrt{\frac{p(1-p)}{n}} \leq \frac{1}{\sqrt{4n}}$, obtendo-se

$$I.C.(p : \gamma) = \left[\hat{p} - z_{\alpha/2} \sqrt{1/4n} ; \hat{p} + z_{\alpha/2} \sqrt{1/4n} \right] \qquad \textbf{(I)}$$

onde z é tal que $Pr(Z < -z) = Pr(Z > z) = \frac{\alpha}{2}$.

O intervalo (I) é chamado *conservativo*, pois se *p* não for igual a 0,5 e estiver próximo de zero ou de um, então, ele fornece um intervalo de amplitude desnecessariamente grande, porque substituímos $p(1-p)$ pelo seu valor máximo 1/4. Assim, a menos que $p = 1/2$, podemos proceder como segue.

Após a retirada de uma amostra piloto, podemos obter um intervalo de confiança para *p*, com um coeficiente de confiança γ qualquer, $0 < \gamma < 1$. Para isso, usamos $\hat{p}(1-\hat{p})$ como estimador de $p(1-p)$. Então, o intervalo fica

$$I.C.(p : \gamma) = \left[\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} ; \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right], \qquad \textbf{(II)}$$

onde z é tal que $Pr(Z < -z) = Pr(Z > z) = \frac{\alpha}{2}$.

9.5. Outros Intervalos de Confiança

9.5.1. Intervalo de Confiança para a Diferença de Duas Médias Populacionais com Variâncias Populacionais Desconhecidas, mas Supostas Iguais.

O intervalo de confiança para $\mu_X - \mu_Y$ com 100γ % de confiança é dado por:

$$I.C.(\mu_X - \mu_Y : \gamma) = \left[(\bar{X} - \bar{Y}) - t_{\alpha/2} . S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}, (\bar{X} - \bar{Y}) + t_{\alpha/2} . S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} \right],$$

com $S_P^2 = \frac{(n_X - 1) S_X^2 + (n_Y - 1) S_Y^2}{n_X + n_Y - 2}$ e $Pr(t_{(n_X + n_y - 2)} < -t) = Pr(t_{(n_X + n_y - 2)} > t) = \frac{\alpha}{2}$.

9.5.2. Intervalo de Confiança para a Diferença de Duas Proporções Populacionais

O intervalo de confiança para $p_1 - p_2$ com 100γ % de confiança, após a retira de amostras piloto, é dado por:

$$I.C.(p_1 - p_2 : \gamma) = \left[(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]$$

com $Pr(Z < -z) = Pr(Z > z) = \frac{\alpha}{2}$.

9.5. Erro de Estimação e Tamanho das Amostras

Acabamos de ver como construir intervalos de confiança para os principais parâmetros populacionais. Em todos os casos, supusemos dado o nível de confiança desses intervalos. Evidentemente, o nível de confiança deve ser fixado de acordo com a probabilidade de acerto que se deseja ter na estimação por intervalo. Sendo conveniente, o nível de confiança pode ser aumentado até tão próximo de 100% quanto se queira, mas isso resultará em intervalos de amplitude cada vez maiores, o que significa perda de precisão na estimação.

É claro que seria desejável termos intervalos com alto nível de confiança e pequena amplitude, o que corresponderia a estimarmos o parâmetro em questão com pequena probabilidade de erro e grande precisão. Isso, porém, requer uma amostra suficientemente grande, pois, para *n* fixo, confiança e precisão variam em sentido opostos.

Veremos a seguir como determinar o erro de estimação e o tamanho das amostras necessárias nos casos de estimação da média ou de uma proporção populacional.

O erro num intervalo de estimação diz respeito a diferença entre a média amostral e a verdadeira média da população. Como o intervalo tem centro na média amostral, o *erro máximo provável* é igual à metade da amplitude do intervalo (semi-amplitude).

Vimos, que o intervalo de confiança para a média μ da população normal quando σ é conhecido tem semi-amplitude dada por:

$$e = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad \textbf{(1)}$$

Fixando *e* e *n* na expressão acima, podemos determinar α, o que equivale a determinar a confiança de um intervalo de amplitude conhecida. Podemos também, fixados α e *e*, determinar *n*, que é o problema da determinação do tamanho da amostra necessária para se realizar a estimação por intervalo com confiança e a precisão desejadas. Deste modo temos que,

$$n = \left(\frac{z_{\alpha/2}\ \sigma}{e} \right)^2 .$$

Está será a expressão usada para a determinação do tamanho da amostra necessária, se σ for conhecido.

Não conhecendo o desvio-padrão da população, deveríamos substituí-lo por sua estimativa S e usar a distribuição t de Student, ou seja, substitui σ por S e usar t de Student na expressão (1). Ocorre, porém, que, não tendo ainda sido retirada a amostra, não dispomos, em geral, do valor de S. Se não conhecemos nem ao menos uma limitação superior para σ, a única solução será, então, colher uma *amostra-piloto* de tamanho *n'* e, com base nela, obtermos uma estimativa S, empregando, a seguir, a expressão

$$n = \left(\frac{t_{n'-1}\ S}{e} \right)^2 .$$

Se $n \leq n'$, a amostra-piloto já terá sido suficiente para a estimação. Caso contrário, deveremos retirar, ainda, da população, os elementos necessários à complementação do tamanho mínimo da amostra.

Procedemos de forma análoga se desejamos estimar uma proporção populacional com determinada confiança e dada precisão. Podemos obter

$$n = \left(\frac{z_{\alpha/2}}{e} \right)^2 . p\,(1 - p) . \qquad \textbf{(2)}$$

O obstáculo à determinação do tamanho da amostra por meio da expressão (2) está em desconhecermos *p* e tampouco dispormos de sua estimativa $\hat{p}$, pois a amostra ainda não foi retirada. Essa dificuldade pode ser resolvida através de uma amostra-piloto, analogamente ao caso descrito na estimação de μ, ou analisando-se o comportamento do fator $p(1-p)$ para $0 \leq p \leq 1$. Pose-se observar facilmente que $p(1-p)$ é a expressão de uma parábola cujo ponto máximo é $p = ½$.

Desse modo, se substituirmos, na expressão (2), $p(1-p)$ por seu valor máximo, ¼, seguramente o tamanho de amostra obtido será suficiente para a estimação, qualquer que seja *p*. Isso eqüivale a considerar

$$n = \left(\frac{z_{\alpha/2}}{e} \right)^2 . \frac{1}{4} = \left(\frac{z_{\alpha/2}}{2e} \right)^2 . \qquad \textbf{(3)}$$

Pelo mesmo raciocínio, se sabemos que seguramente $p \leq p_0 \leq ½$ ou $p \geq p_0 \geq ½$, podemos usar o limite p_0 ao invés de p, na expressão (2), obtendo um tamanho de amostra suficiente, pois teremos então $p(1-p) \leq p_0(1-p_0)$.

Evidentemente, usando-se a expressão (3), corre-se o risco de dimensionar uma amostra bem maior do que a realmente necessária. Isso ocorrerá se p for, na realidade, próximo de 0 ou 1. Se o custo envolvido for elevado e proporcional ao tamanho da amostra, será desejável evitar que tal fato ocorra, sendo mais prudente a tomada de uma amostra-piloto. Inversamente, em muitos casos, é preferível, por simplificação, proceder conforme indicado, com base em uma limitação superior para o fator $p(1-p)$.

Exercícios

Exercício 1: O peso médio ao nascer e o desvio padrão de bezerros da raça Ibagé, examinada uma amostra de 20 partos, foram de 26 Kg e 2 Kg, respectivamente. Dê uma estimativa por intervalo do verdadeiro peso médio, utilizando grau de confiança de 95%. [25,064; 26,936]

Exercício 2: Coletou-se uma amostra de 35 peixes da espécie Xenomelaniris brasiliensis, na localidade praia da Barra da Lagoa, Florianópolis, SC, a qual apresentou 45,7% de peixes com comprimento total acima de 50 mm. Encontre um intervalo, com 99% de confiança, dentro do qual deve estar a verdadeira proporção de peixes dessa espécie com comprimento acima de 50 mm. [0,240; 0,674]

Exercício 3: Com relação ao exercício 1, que tamanho de amostra será necessário para produzir um intervalo de confiança de 95% para a verdadeira média, com uma precisão de 2% da média? (aproximadamente 65)

Exercício .4: Num cultivo de Crassostrea gigas, conhecida popularmente como Ostra do Pacífico, na localidade da Praia da Pinheira, Palhoça, SC, foram semeadas duas mil duzentas e vinte sementes desses ostreídeos, chegando-se a uma população adulta de duas mil ostras (N=2000). Por meio de trabalhos similares, sabe-se que o desvio padrão populacional do peso da parte mole, em gramas, vale 8 g. Com o objetivo de estimar o peso médio da parte mole, qual deverá ser o tamanho da amostra necessário para que o erro amostral seja de no máximo igual a 1 g, para mais ou para menos, com grau de confiança de 95%? (aproximadamente 246)

Exercício 5: Dos 10000 bovinos de uma raça retira-se um lote de 64 animais, e obtém-se o peso médio de 260 Kg e desvio padrão de 16 Kg.

(a) Quais os limites de confiança a 95% para o peso médio populacional? ([256,08; 263,92])

(b) Com que confiança dir-se-ia que o peso médio é $260 \pm 0{,}834$? (32,56%)

(c) Que tamanho deve ter a amostra para que seja de 95% a confiança na estimativa $260 \pm 0{,}79932$? (1540)