Estatística Aplicada a Recursos Hídricos

Docente: Rachid Muleia

rachid.muleia@uem.mz

Mestrado em Gestão de Recursos Hídricos - DGEO/UEM

Tema: Inferência Estatística: Intervalo de Confiança

Ano lectivo: 2023

Intervalo de Confiança: Motivação

- Os estimadores pontuais fornecem como estimativa um único valor numérico para o parâmetro de interesse, mas não fornecem qualquer informação sobre a precisão e a confiabilidade dessa estimativa;
- Ou seja, não se sabe quão próximo a estimativa pode estar do verdadeiro valor do parâmetro.
- Uma alternativa para apresentar um único valor para o parâmetro é calcula-lo e estabelecer um intervalo completo de valores plausíveis;
- Ou seja, uma estimativa de intervalo ou "Intervalo de confiança" (IC)

Intervalo de Confiança

- Intervalo de Confiança (IC) fornece uma estimativa mais informativa para o parâmetro de interesse, que inclui uma medida de precisão do valor obtido;
- O IC incorpora, à estimativa pontual do parâmetro, informação a respeito da variabilidade do estimador.
- Os IC são obtidos através da distribuição amostral de seus estimadores;
- Aqui, serão abordados IC para média (μ) e para proporção (p) da população.

IC para μ de uma população Normal, com σ^2 conhecida

- Para construir IC para μ, recorremos às proposições sobre distribuição amostral da média, $\bar{X}$.

- Considere uma amostra aleatória de tamanho n, $X_1, X_2, \ldots, X_n$, de uma distribuição normal com média μ e variância σ^2. Então

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{Padronizando o } \bar{X}, \quad Z = \frac{\bar{X} - \mu}{\frac{\sigma}{n}} \sim N(0,1)$$

Observações:

- Sabe-se que o conhecimento do centro da população precede tipicamente informações relacionadas à dispersão.

- Significa que, se o valor de μ é desconhecido, é improvável que o valor de σ esteja disponível

- ou seja, a suposição sobre o conhecimento de σ^2 aqui adoptada não razoável.

IC para μ de uma população Normal, com σ^2 conhecida

Procedimento para a construção do IC:

i) Consideramos uma amostra aleatória de n elementos;

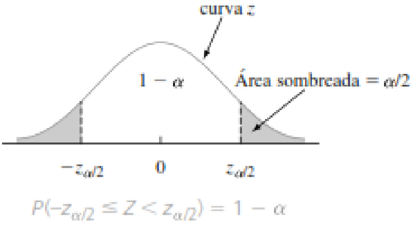
ii) Calculamos a média amostral $\bar{x}$.

iii) Calculamos o desvio padrão da média amostral: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

iv) Fixamos o nível de significância α, e com ele determina-se $z_{1-\frac{\alpha}{2}}$, tal que

$$P\left(|Z| < z_{1-\frac{\alpha}{2}}\right) = P\left(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

IC para μ de uma população Normal, com σ^2 conhecida



O valor de $z_{1-\frac{\alpha}{2}}$ pode ser obtido da tabela Normal padrão. Assim

$$P\left(z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = P\left(z_{1-\frac{\alpha}{2}} < \frac{\bar{x}-\mu}{\frac{\sigma}{n}} < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Assim, o IC para μ com nível de confiança $1-\alpha$ é dado por:

$$IC(\mu,(1-\alpha)\%) = \left[\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

Interpretação do IC

- Sabe-se que, antes que uma amostra seja colectada, $\bar{X}$ é uma v.a. e, por tanto, o IC obtido também é aleatório;

- Nesse caso, a probabilidade de que IC contenha o verdadeiro valor de μ é dada por $1-\alpha$

- No entanto, ao colectar a amostra, $\bar{x}$ torna-se $\bar{x}_{obs}$, e como conhecemos σ, n e $z_{1-\frac{\alpha}{2}}$, o IC passa a ser numérico;

- Ou seja, o IC não é aleatório e μ é constante (desconhecido), de modo que seria errado afirmar $P(\mu \in IC(\mu,(1-\alpha)\%)) = 1-\alpha$;

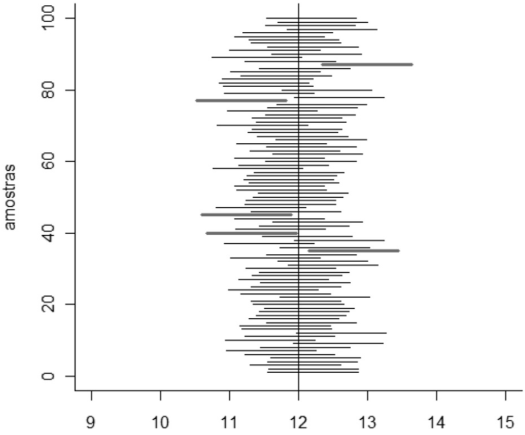
- Com intervalo construído, μ pode estar dentro dele ou não. Não há mais qualquer probabilidade envolvida.

- A interpretação conveniente do IC é baseada na frequência relativa (abordagem frequencista)

- Dizer que um evento $A = \mu \in IC(\mu,(1-\alpha)\%)$ tem probabilidade de 0,95 significa dizer que, se o experimento no qual A é definido for repetido várias vezes, em longo prazo, A ocorrerá 95% das vezes.

Interpretação do IC

Interpretação do IC: Se obtivermos várias amostras de mesmo tamanho *n* e, para cada uma delas, calcularmos os correspondentes IC com nível de confiança de $(1-\alpha)\%$, esperamos que a proporção de intervalos que contenham o verdadeiro valor de μ seja igual a $1-\alpha$;



Interpretação do IC

- A amplitude do IC é igual a $2z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$
- $z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ é designado o erro de estimação.
- Quanto maior o valor de $z_{1-\frac{\alpha}{2}}$ maior é a amplitude do IC.
- Se imaginarmos a amplitude do IC como a especificação de sua precisão ou acurácia, então seu nível de confiança (a confiabilidade) estará inversamente relacionado a sua precisão;
- O IC com maior grau de certeza em relação ao valor da média pode ser impreciso quando seus limites extremos estiverem muito distantes;
- Por outro lado, se desejamos tornar um IC menor sem reduzir o nível de confiança, precisamos seleccionar amostra maior;
- Conforme o tamanho n da amostra aumenta, o erro padrão $\sigma/\sqrt{n}$ diminui, o que resulta em IC mais estreito (preciso).

Construção de um IC

Exemplo 1: Suponha que a altura de atletas que praticam um certo tipo de modalidade siga o modelo normal com média μ desconhecida e variância igual a $0,01\ m^2$. Uma amostra de 10 atletas foi sorteada e forneceu média de $1,69\ m$. Determine um intervalo de 95% de confiança para o verdadeiro valor da média μ.

Construção de um IC

Exemplo 1: Suponha que a altura de atletas que praticam um certo tipo de modalidade siga o modelo normal com média μ desconhecida e variância igual a $0,01\ m^2$. Uma amostra de 10 atletas foi sorteada e forneceu média de $1,69\ m$. Determine um intervalo de 95% de confiança para o verdadeiro valor da média μ.

Resposta: Seja X_i "Altura do atleta i". Como $X_i \sim N(\mu, 0,01) \Rightarrow$
$\bar{X} \sim N\left(\mu, \frac{0,01}{10}\right)$. Para $\alpha = 0,05$, tem-se $z_{1-\frac{\alpha}{2}} = 1,96$. Logo o intervalo de confiança de 95% para μ é

$$IC(\mu, 95\%) = \left[1,69 - 1,96\sqrt{\frac{0,01}{10}}; 1,69 + 1,96\sqrt{\frac{0,01}{10}}\right]$$
$$= [1,63; 1,75]. \ \blacksquare$$

Interpretação: Se extrairmos 100 amostras, cada uma com 10 animais, e para cada uma delas, construirmos IC de 95%, aplicando o mesmo procedimento, espera-se que 95 desses intervalos contenham o verdadeiro valor da média μ $\blacksquare$.

Construção de um IC

Exemplo 2: Um provedor de acesso à Internet está a monitorar a duração do tempo das conexões de seus clientes, com o objectivo de dimensionar seus equipamentos. São desconhecidas a média e a distribuição de probabilidade desse tempo, mas o desvio padrão, por analogia a outros serviços, é considerado igual $\sqrt{50}$ a minutos. Uma amostra de 500 conexões resultou num valor médio observado de 25 minutos. O que dizer da verdadeira média, com confiança 95%?

Construção de um IC

Exemplo 2: Um provedor de acesso à Internet está a monitorar a duração do tempo das conexões de seus clientes, com o objectivo de dimensionar seus equipamentos. São desconhecidas a média e a distribuição de probabilidade desse tempo, mas o desvio padrão, por analogia a outros serviços, é considerado igual $\sqrt{50}$ a minutos. Uma amostra de 500 conexões resultou num valor médio observado de 25 minutos. O que dizer da verdadeira média, com confiança 95%?

Resposta: Seja X_i "O tempo de duração da conexão i". A distribuição de X é desconhecida. No entanto, como $n = 500$, tamanho da amostra, é relativamente grande, então é razoável aplicar o Teorema Central do Limite. Portanto, $\bar{X} \sim N\left(\mu, \frac{50}{500}\right)$. Para $(1-\alpha)\% = 95$, $z_{1-\frac{\alpha}{2}} = 1,96$. Então

$$IC(\mu, 95\%) = \left[25 - 1,96\sqrt{\frac{50}{500}}; 25 + 1,96\sqrt{\frac{50}{500}}\right]$$
$$= [24,38; 25,62]. \blacksquare$$

IC para amostras grandes

- Na prática consideramos uma amostra grande quando $n > 30$
- Aqui, o objectivo é construir IC para parâmetros de populações não normais, com distribuições aproximadamente normais, ou então de populações normais com variâncias desconhecidas

IC para proporção da população p

Seja p a proporção de "'sucessos' de uma população em que sucesso identifica indivíduo ou objecto que tenha uma característica especificada. Uma amostra aleatória de n indivíduos será seleccionada e X será o número de sucessos na amostra.

- Nesse caso, sabe-se que, $X \sim Bin(n, p)$, onde $E(X) = np$ e $\sigma_X = \sqrt{np(1-p)}$
- Além disso, se n for "suficientemente grande", se ambos np e $n(1-p)$ forem maiores ou iguais a 5 (algumas pessoas acreditam na condição mais conservadora:$np \geq 10$ e $n(1-p) \geq 10$), X possui aproximadamente uma distribuição normal.

IC para proporção da população

Sabe-se que o estimador natural de p é $\hat{p} = \frac{X}{n}$. Uma vez que $\hat{p}$ é apenas X multiplicado pela constante $1/n$, então $\hat{p}$ também tem distribuição aproximadamente normal com

$$
\begin{aligned}
E(\hat{p}) &= E\left(\frac{1}{n}X\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p \\
Var(\hat{p}) &= Var\left(\frac{1}{n}X\right) = \frac{1}{n^2}Var(X) = \frac{p(1-p)}{n}
\end{aligned}
$$

Logo $\hat{p} \cong N\left(p, \frac{p(1-p)}{n}\right)$ e $\frac{\hat{p}-p}{\sigma_{\hat{p}}} \cong N(0,1)$.

$$
P\left(-z_{1-\frac{\alpha}{2}} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha
$$

O IC para p é então dado por:

$$
IC(p,(1-\alpha)\%) = \left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}; \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right]
$$

onde $\hat{q} = 1 - \hat{p}$

IC para proporção da população

Exemplo 3: Considere a sobrevivência de cinco anos entre os pacientes diagnosticados com câncro nos pulmões. A proporção média de indivíduos que sobrevive é $p = 0,10$; o desvio padrão é $\sqrt{p(1-p)} = \sqrt{0,10 \cdot 0,90} = 0,30$. Se seleccionarmos amostras repetidas de tamanho 50 dessa população, que fracção terá proporção amostral $\hat{p} = 0,20$ ou mais?

IC para proporção da população

Exemplo 3: Considere a sobrevivência de cinco anos entre os pacientes diagnosticados com câncro nos pulmões. A proporção média de indivíduos que sobrevive é $p = 0,10$; o desvio padrão é $\sqrt{p(1-p)} = \sqrt{0,10 \cdot 0,90} = 0,30$. Se seleccionarmos amostras repetidas de tamanho 50 dessa população, que fracção terá proporção amostral $\hat{p} = 0,20$ ou mais?

Resposta: Uma vez que $np = 50(0,10) = 5$ e $n(1-p) = 50(0,90) = 45$, com esse tamanho da amostra, é razoável aplicar o TCL, ou seja,

$$\hat{p} \cong N\left(p, \frac{p(1-p)}{n}\right) \quad \text{e} \quad Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

$p = 0,10$ e $\sqrt{p(1-p)/n} = \sqrt{0,10(0,9)/50} = 0,0424$. Portanto

$$\begin{aligned} P(\hat{p} \geq 0,20) =& P\left(Z \geq \frac{0,20 - 0,10}{0,0424}\right) = P(Z \geq 2,36) \\ =& 1 - \Phi(2,36) = 0,009 \end{aligned}$$

Somente cerca de $0,9\%$ das amostras terá uma proporção amostral de $0,20$ ou mais.

IC para proporção da população

Exemplo 4: Para se estimar a percentagem de alunos de um curso favoráveis à modificação do currículo escolar, tomou-se uma amostra de 100 alunos, dos quais 80 foram favoráveis.

i) Construa um IC para a proporção de todos os alunos do curso favoráveis à modificação ao nível nível de significância 4%.

ii) Qual o valor do erro de estimação cometido em a)?

IC para proporção da população

Exemplo 4: Para se estimar a percentagem de alunos de um curso favoráveis à modificação do currículo escolar, tomou-se uma amostra de 100 alunos, dos quais 80 foram favoráveis.

i) Construa um IC para a proporção de todos os alunos do curso favoráveis à modificação ao nível nível de significância 4%.

ii) Qual o valor do erro de estimação cometido em a)?

Resolução: Seja X: "nº de alunos favoráveis à modificação"; $x = 80$; $\alpha = 4\%$;

$$\hat{p} = \frac{x}{n} = \frac{80}{100} = 0,80 \text{ e } \hat{q} = 0,20$$

$$\sigma_{\hat{p}} \cong \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0,80 \cdot 0,20}{100}} = 0,04$$

Como $\alpha = 0,04$, então $z_{1-\frac{\alpha}{2}} = 2,05$

IC para proporção da população

Resolução Exemplo 4:

i) IC para a proporção ao nível nível de significância 4%:

$$
\begin{aligned}
IC(p, 96\%) =& [0,8 - 2,05 \cdot 0,04; 0,8 + 2,05 \cdot 0,04] \\
=& [0,7180; 0,882] \\
=& [71,80\%; 88,2\%]
\end{aligned}
$$

ii) O valor do erro de estimação cometido em a):

$$
\begin{aligned}
z_{1-\frac{\alpha}{2}} =& \frac{\hat{p} - p}{\sigma_{\hat{p}}} \rightarrow z_{1-\frac{\alpha}{2}} = \frac{e}{\sigma_{\hat{p}}} \;\therefore e = z_{1-\frac{\alpha}{2}} \cdot \sigma_{\hat{p}} \\
e =& 2,05 \cdot 0,04 = 0,082 \;\therefore e = 8,2\%
\end{aligned}
$$

IC para proporção da população

Exemplo 5: Considere a distribuição de sobrevivência de cinco anos para os indivíduos abaixo dos 40 anos diagnosticados com câncro nos pulmões. Essa distribuição tem média da população p desconhecida. Em uma amostra aleatoriamente seleccionada de 52 pacientes, somente seis sobreviveram cinco anos. Construa o IC para p ao nível de confiança de 95%.

IC para proporção da população

Exemplo 5: Considere a distribuição de sobrevivência de cinco anos para os indivíduos abaixo dos 40 anos diagnosticados com câncro nos pulmões. Essa distribuição tem média da população p desconhecida. Em uma amostra aleatoriamente seleccionada de 52 pacientes, somente seis sobreviveram cinco anos. Construa o IC para p ao nível de confiança de 95%.

Resolução: Seja X: "nº de pacientes que sobreviveram cinco anos"; $x = 6$; $\alpha = 5\%$;

$$\hat{p} = \frac{x}{n} = \frac{6}{52} = 0,115 \text{ e } \hat{q} = 0,885$$

Como $n\hat{p} = 52(0,115) = 6,0$ e $n(1-\hat{p}) = 52(0,885) = 46,0$, justifica o uso da aproximação normal. Portanto $\sigma_{\hat{p}} = \sqrt{\frac{0,115 \cdot 0,885}{52}} = 0,044$,

$$\begin{aligned} IC(p, 95\%) =& [0,115 - 1,96 \cdot 0,044; 0,115 + 1,96 \cdot 0,044] \\ =& [0,028; 0,202] = [2,80\%; 20,20\%] \end{aligned}$$

Estamos 95% confiantes de que esse intervalo contenha a proporção verdadeira dos indivíduos que sobrevivem cinco anos.

IC para μ de uma população Normal, com σ^2 desconhecida

Consideramos uma situação em que a população de interesse é normal com μ e σ^2 desconhecidos,

- Se n for suficientemente grande ($n > 30$), s^2 (variância amostral) se aproxima bastante de σ^2. Nesse caso, a v.a.

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

possui aproximadamente uma distribuição normal padrão. Isso implica que

$$IC(\mu, (1-\alpha)\%) = \left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}; \bar{x} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right]$$

Essa fórmula é válida independentemente do formato da distribuição da população.

IC para μ de uma população Normal, com σ^2 desconhecida

- Quando n é pequeno ($n \leq 30$), a razão $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ não tem distribuição normal padrão, pois além da v.a. $\bar{X}$ no numerador, há também variabilidade de s no denominador, de modo que a distribuição de probabilidade dessa razão é mais dispersa que a normal padrão.

- Nesse caso, a v.a.

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

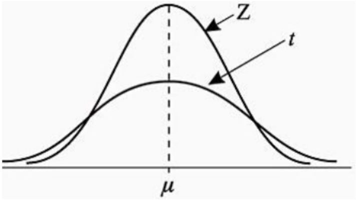
  possui uma distribuição t de Student com $n-1$ graus de liberdade (gl). Denotamos essa representação por t_{n-1}.

Principais propriedades das distribuições t de Student

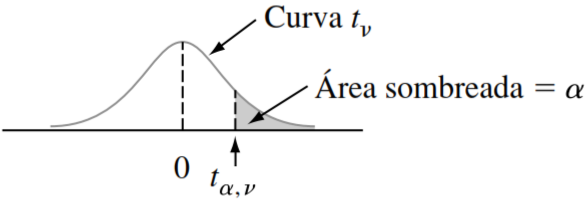
Seja t_{n-1} a curva da função densidade associada a $n-1$ gl: a curva t_{n-1} é simétrica e está centrada em $\mu = 0$; toda curva t_{n-1} é mais dispersa que a curva normal padrão (z); à medida que $n-1 \to \infty$, a dispersão da t_{n-1} correspondente diminui; à medida que $n-1 \to \infty$, a sequência das curvas t_{n-1} se aproxima da curva $N(0,1)$

IC para μ de uma população Normal, com σ^2 desconhecida

Gráfico comparativo entre a distribuição t a Z:



Uso da tabela para a obtenção de valores críticos para a distribuição t, denotados por $t_{n-1,\alpha}$.



IC para μ de uma população Normal, com σ^2 desconhecida

Exemplo: Dados n e $\alpha = 5\% \rightarrow P(t > t_{n-1,\frac{\alpha}{2}}) = 0,05$. Encontre $t_{n-1,\frac{\alpha}{2}}$ para $n = 15, 49$ e 121.

			α		
ν	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
50	1,299	1,676	2,009	2,403	2,678
60	1,296	1,671	2,000	2,390	2,660
120	1,289	1,658	1,980	2,358	2,617
∞	1,282	1,645	1,960	2,326	2,576

Resposta: $t_{14;0,025} = 2,145$; $t_{50;0,025} = 2,009$; $t_{120;0,025} = 1,980$.

IC para μ de uma população Normal, com σ^2 desconhecida

Exemplo: Uma amostra constituída de 12 medidas de tensão de ruptura de um fio de algodão apresentou média de $7,38\ kg$ e desvio padrão de $1,24\ kg$. Determinar os intervalos de confiança de 95% e 99% para a média da população.

IC para μ de uma população Normal, com σ^2 desconhecida

Exemplo: Uma amostra constituída de 12 medidas de tensão de ruptura de um fio de algodão apresentou média de $7,38\ kg$ e desvio padrão de $1,24\ kg$. Determinar os intervalos de confiança de 95% e 99% para a média da população.

Resposta: $n = 12$, $\bar{x} = 7,38\ kg$, $s = 1,24\ kg$, então $s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{1,24}{\sqrt{12}} = 0,358$; $gl = 12 - 1 = 11$; $t_{n;\frac{\alpha}{2}} = t_{11;0,025} = 2,201$; $t_{11;0,005} = 3,106$.

Os intervalos de confiança são definidos da seguinte forma

$$
\begin{aligned}
IC(\mu, 95\%) =& [7,38 - 2,201 \cdot 0,358; 7,38 + 2,201 \cdot 0,358] \\
=& [6,592; 8,168] \\
IC(\mu, 99\%) =& [7,38 - 3,106 \cdot 0,358; 7,38 + 3,106 \cdot 0,358] \\
=& [6,268; 8,492]
\end{aligned}
$$