Estatística Aplicada a Recursos Hídricos

Docente: Rachid Muleia

rachid.muleia@uem.mz

Mestrado em Gestão de Recursos Hídricos - DGEO/UEM

Tema: Inferência Estatística: Estimação Pontual

Ano lectivo: 2023

Conceitos básicos em inferência estatística

- A Inferência Estatística fornece um conjunto de técnicas que objectiva estudar a população através de evidências fornecidas por uma amostra.
- É a amostra que contém os elementos que podem ser observados e, a partir daí, afirmações sobre quantidades de interesse podem ser feitas.
- O problema fundamental da inferência estatística, é medir o grau de incerteza ou risco das generalizações sobre a população a partir das conclusões baseadas nos resultados da amostra.

Conceitos básicos em inferência estatística

- População: é o conjunto formado por indivíduos ou objectos que têm pelo menos uma característica (variável) comum e observável. Por exemplo:
  - população dos alunos do primeiro ano de uma faculdade;
  - população de peças fabricadas numa linha de produção.
- Amostra: Qualquer subconjunto formado exclusivamente por elementos de uma certa população. Detona-se por n o número de elementos da amostra, o seu tamanho.
- Amostragem: é o processo de selecção de uma amostra, que possibilita o estudo das características da população. A generalização dos resultados só é valida quando a amostra seleccionada é representativa da população em estudo.
- Erro amostral: é o erro que ocorre justamente pelo uso da amostra.

Conceitos básicos em inferência estatística

- Parâmetro: é a medida usada para descrever uma característica numérica populacional. É uma quantidade da população, em geral desconhecida, sobre a qual interessa estudar. Genericamente representaremos por θ. Por exemplo: a média (μ) e a variância (σ^2).
- Estimador: É uma característica numérica determinada na amostra, em função de seus elementos, com a finalidade de representar, ou estimar, um parâmetro de interesse na população. Em geral, representa-se por $\hat{\theta}$. Por exemplo: a média amostral ($\bar{x}$) e variância amostral (s^2) são estimadores de μ e σ^2, respectivamente.
- Estimativa: é o valor numérico determinado pelo estimador. Geralmente, denota-se por θ_0.
- Logo o erro amostral, denotado por ε, é definido por $\varepsilon = \hat{\theta} - \theta$

Conceitos básicos em inferência estatística

Resumo: Denote os parâmetros média, variância e proporção de certa característica na população por μ, σ^2 e p, respectivamente. Os estimadores "naturais" para estas quantidades são as correspondentes média, variância e proporção calculadas na amostra. Representando-os, respectivamente, por $\bar{X}$, $\hat{\sigma}^2$ e $\hat{p}$, temos

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \sum_{i=1}^{n} \frac{X_i}{n};$$

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2;$$

$$\hat{p} = \frac{\text{número de itens com a característica na amostra}}{n}$$

- Note que cada um dos estimadores, digamos $\hat{\theta}_i$ é uma função das variáveis aleatórias constituintes da amostra, isto é, $\hat{\theta}_i = f(X_1, X_2, \ldots, X_n)$. Logo, um estimador também é uma variável aleatória.

Conceitos básicos em inferência estatística

Exemplo: Para estudar o nível de colesterol em uma população de atletas, colectou-se uma amostra de 10 jovens atletas, obtendo os seguintes valores: $180, 196, 185, 165, 190, 195, 180, 176, 165$ e 195.

- Suponha que o interesse seja o nível médio de colesterol de toda população (na qual não se tem pleno acesso). Então estima-se o parâmetro μ (desconhecido) pela média amostral ($\bar{X}$) calculada com os valores da amostra:

Conceitos básicos em inferência estatística

Exemplo: Para estudar o nível de colesterol em uma população de atletas, colectou-se uma amostra de 10 jovens atletas, obtendo os seguintes valores: $180, 196, 185, 165, 190, 195, 180, 176, 165$ e 195.

- Suponha que o interesse seja o nível médio de colesterol de toda população (na qual não se tem pleno acesso). Então estima-se o parâmetro μ (desconhecido) pela média amostral ($\bar{X}$) calculada com os valores da amostra:

$$\bar{x}_{obs} = \frac{180 + 196 + 185 + \ldots + 176 + 165 + 195}{10} = 182,7$$

- Considere que para a população em estudo (de jovens atletas), classifiquemos como tendo "taxa alta" os atletas com valores acima de 190 e "taxa baixa", os demais. Sendo X_i o nível de colesterol do i-ésimo atleta escolhido:

$$Y_i = \begin{cases} 1, & \text{se } X_i > 190; \\ 0, & \text{se } X \leq 190. \end{cases}$$

  Assim, Y_i é uma v.a. com distribuição de Bernoulli, que assume 1 para taxas altas e 0 para as baixas.

Conceitos básicos em inferência estatística

Exemplo (Cont.): Assim, obtém-se a seguinte tabela:

i	1	2	3	4	5	6	7	8	9	10
X_i	180	196	185	165	190	195	180	176	165	195
Y_i	0	1	0	0	0	1	0	0	0	1

A proporção p de atletas com taxa de colesterol alta será estimada pela proporção de taxas altas encontradas na amostra, $\hat{p}$, ou seja:

$$\hat{p}_{obs} = \frac{Y_1 + Y_2 + \ldots + Y_{10}}{10} = \frac{0 + 1 + \ldots + 1}{10} = 0,3$$

Portanto, Com base nessa amostra, pode se assumir que 30% de todos os atletas têm taxa relativamente alta de colesterol.

- Agora suponha que se deseja estudar a variabilidade de X. Consideramos como parâmetro de interesse a variância (σ^2), cujo estimador é:

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\hat{\sigma}^2_{obs} = \frac{1}{10}\left[(180 - 182,7)^2 + \ldots + (195 - 182,7)^2\right] = 136,0111$$

Propriedades de um estimador

As principais propriedades de um estimador são:

i) Estimador não viciado

Um estimador $\hat{\theta}$ é não viciado ou não enviesado ou não tendencioso para um parâmetro θ se $E(\hat{\theta}) = \theta$ 0. Em outras palavras, um estimador é não viciado se o seu valor esperado coincide com o parâmetro de interesse.

Exemplo: Se $X \sim N(\mu, \sigma^2)$, então $\bar{x}$ é um estimador não viciado de μ e $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ é estimador não viciado de σ^2.

Demonstração:

Propriedades de um estimador

As principais propriedades de um estimador são:

i) Estimador não viciado

Um estimador $\hat{\theta}$ é não viciado ou <u>não enviesado</u> ou <u>não tendencioso</u> para um parâmetro θ se $E(\hat{\theta}) = \theta$ 0. Em outras palavras, um estimador é não viciado se o seu valor esperado coincide com o parâmetro de interesse.

Exemplo: Se $X \sim N(\mu, \sigma^2)$, então $\bar{x}$ é um estimador não viciado de μ e $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ é estimador não viciado de σ^2.

Demonstração:

$$
\begin{aligned}
E(\bar{x}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{n} E(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\left[E(x_1) + E(x_2) + \ldots + E(x_n)\right] \\
&= \frac{1}{n} \cdot n\mu = \mu
\end{aligned}
$$

Logo, $\bar{x}$ é um estimador não tendencioso de μ.

Propriedades de um estimador

Demonstração (Estimador não viciado)

$$
\begin{aligned}
E(s^2) =& E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] \\
=& \frac{1}{n-1}E\left[\sum_{i=1}^{n}(x_i^2-2x_i\cdot\bar{x}+\bar{x}^2)\right] \\
=& \frac{1}{n-1}E\left[\sum_{i=1}^{n}x_i^2-2\cdot\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2\right] \\
=& \frac{1}{n-1}E\left[\sum_{i=1}^{n}x_i^2-2\cdot\bar{x}\cdot n\cdot\frac{\sum_{i=1}^{n}x_i}{n}+n\bar{x}^2\right]
\end{aligned}
$$

Propriedades de um estimador

Demonstração (Estimador não viciado)

$$E(s^2) = \frac{1}{n-1} E\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(x_i^2) - nE(\bar{x}^2)\right]$$

Considerando que $E(x^2) = \sigma^2 + \mu^2$ e $E(\bar{x}^2) = \frac{\sigma^2}{n} + \mu^2$, tem-se

$$\begin{aligned} E(s^2) =& \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] = \frac{1}{n-1}\left[n\sigma^2 - \sigma^2\right] \\ =& \frac{1}{n-1}[\sigma^2(n-1)] = \sigma^2 \end{aligned}$$

o que demonstra que s^2 é um estimador não viciado de σ^2

Propriedades de um estimador

ii) Estimador Consistente

Um estimador $\hat{\theta}$ é consistente, se, à medida que o tamanho da amostra (n) aumenta, seu valor esperado converge para o parâmetro de interesse (θ) e sua variância converge para zero. Ou seja, $\hat{\theta}$ é consistente se satisfaz as propriedades:

$$i) \lim_{n\to\infty} E(\hat{\theta}) = \theta;$$

$$ii) \lim_{n\to\infty} Var(\hat{\theta}) = 0;$$

A média amostral ($\bar{X}$) é um estimador consistente, uma vez que:

$$\begin{aligned} E(\bar{X}) =& E\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{n\mu}{n} = \mu \\ Var(\bar{X}) =& Var\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned}$$

Portanto, já que $Var(\bar{X}) = \sigma^2/n \to 0$ conforme $n \to \infty$, então $\overline{X}$ é consistente para μ.

Propriedades de um estimador

iii) Estimador Eficiente

Dados dois estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$, não viciados para um parâmetro θ, dizemos que $\hat{\theta}_1$ é mais eficiente do que $\hat{\theta}_2$ se $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$.

A distribuição da média amostral

- A importância da média amostral $\bar{X}$ surge de seu uso para tirar conclusões sobre a média da população μ desconhecida;

- A maioria dos procedimentos de inferência baseia-se nas propriedades da distribuição de $\bar{X}$.

- Preposição: Sejam $X_1, X_2, \ldots, X_n$ elementos da amostra aleatória de uma distribuição com valor médio μ e desvio padrão σ. Então,

$$1 - E(\bar{X}) = \mu$$

$$2 - Var(\bar{X}) = \frac{\sigma^2}{n} \text{ e } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Além disso, com $T_0 = X_1 + X_2 + \ldots + X_n$ (O total da amostra), $E(T_0) = n\mu$ e $Var(T_0) = n\sigma^2$ e $\sigma_{T_0} = \sqrt{n}\sigma$.

A distribuição da média amostral

- Preposição (Caso de Distribuição da População Normal): Sejam $X_1, X_2, \ldots, X_n$ elementos da amostra aleatória de uma distribuição normal com média μ e desvio padrão σ. Então, para qualquer n, $\bar{X}$ é normalmente distribuído (com média μ e desvio padrão $\sigma/\sqrt{n}$), como é T_0 (com média $n\mu$ e desvio padrão $\sqrt{n}\sigma$

Exemplo: Suponha que a aceitação de um lote de 1000 peças ocorra apenas, se o comprimento médio de 10 peças, retiradas aleatoriamente do lote, estiver entre 5 e 10 cm. Sabe-se que o comprimento das peças é uma variável aleatória com distribuição Normal de média $7,5$ cm e variância 20 cm^2. Qual é a probabilidade de aceitação do lote?

A distribuição da média amostral

- Preposição (Caso de Distribuição da População Normal): Sejam $X_1, X_2, \ldots, X_n$ elementos da amostra aleatória de uma distribuição normal com média μ e desvio padrão σ. Então, para qualquer n, $\bar{X}$ é normalmente distribuído (com média μ e desvio padrão $\sigma/\sqrt{n}$), como é T_0 (com média $n\mu$ e desvio padrão $\sqrt{n}\sigma$

Exemplo: Suponha que a aceitação de um lote de 1000 peças ocorra apenas, se o comprimento médio de 10 peças, retiradas aleatoriamente do lote, estiver entre 5 e 10 cm. Sabe-se que o comprimento das peças é uma variável aleatória com distribuição Normal de média $7,5$ cm e variância 20 cm^2. Qual é a probabilidade de aceitação do lote?

Resposta: Seja X_i o comprimento da i-ésima peça retirada, $i = 1, \ldots, 10$. Temos que a média das 10 peças $\bar{X}$ tem distribuição normal com média $\mu = 7,5$ cm e variância $\sigma^2_{\bar{X}} = 20/10 = 2$. Logo a probabilidade

$$\begin{aligned} P(5 < \bar{X} < 10) =& P\left(\frac{5-7,5}{\sqrt{2}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{10-7,5}{\sqrt{2}}\right) \\ =& P(-1,77 < Z < 1,77) = 0,9232. \ \blacksquare \end{aligned}$$

A distribuição da média amostral

Exemplo: O tempo que um rato de determinada subespécie, seleccionado aleatoriamente leva para encontrar o caminho em um labirinto é uma v.a. distribuída normalmente com $\mu = 1,5$ *min* e $\sigma = 0,35$ *min*. Suponha que cinco ratos sejam seleccionados. Sejam $X_1, \ldots, X_5$ seu tempo no labirinto. Assumindo que X_i, $i = 1, \ldots, 5$ é uma amostra aleatória dessa distribuição normal, qual é a probabilidade de o tempo total $T_0 = X_1 + \ldots + X_5$ dos cinco estar entre 6 e 8 min? e $P(\bar{X} \leq 2) =$?

A distribuição da média amostral

Exemplo: O tempo que um rato de determinada subespécie, seleccionado aleatoriamente leva para encontrar o caminho em um labirinto é uma v.a. distribuída normalmente com $\mu = 1,5$ *min* e $\sigma = 0,35$ *min*. Suponha que cinco ratos sejam seleccionados. Sejam $X_1, \ldots, X_5$ seu tempo no labirinto. Assumindo que X_i, $i = 1, \ldots, 5$ é uma amostra aleatória dessa distribuição normal, qual é a probabilidade de o tempo total $T_0 = X_1 + \ldots + X_5$ dos cinco estar entre 6 e 8 min? e $P(\bar{X} \leq 2) =?$

Resposta: Pela proposição, T_0 possui distribuição normal com $\mu_{T_0} = n\mu = 5(1,5) = 7,5$ e variância $\sigma^2_{T_0} = n\sigma^2 = 5(0,1225) = 0,6125$, assim, $\sigma_{T_0} = 0,783$. Além disso, $\mu_{\bar{X}} = \mu = 1,5$ e $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 0,35/\sqrt{5} = 0,1565$. Logo a probabilidade

$$\begin{aligned} P(6 < T_0 < 8) =& P\left(\frac{6-7,5}{0,783} \leq Z \leq \frac{8-7,5}{0,783}\right) = P(-1,92 \leq Z \leq 0,64) \\ =& \Phi(0,64) - \Phi(-1,92) = 0,7115. \blacksquare \\ P(\bar{X} \leq 2,0) =& P\left(Z \leq \frac{2,0-1,5}{0,1565}\right) = \Phi(3,19) = 0,9993 \blacksquare \end{aligned}$$

Teorema do Limite Central

- Quando os X_i's são distribuídos normalmente, $\bar{X}$ também é para cada tamanho de amostra n.

- Uma hipótese razoável é que, se n for grande, uma curva normal adequada aproximará a distribuição real de $\bar{X}$

Teorema do Limite Central: $X_1, X_2, \ldots, X_n$ formam a amostra aleatória de uma distribuição com média μ e variância σ^2. Então, se n é suficientemente grande, $\bar{X}$ tem aproximadamente uma distribuição normal $\mu_{\bar{x}} = \mu$ e $\sigma^2_{\bar{X}} = \sigma^2/n$, e T_0 também tem aproximadamente uma distribuição normal com $\mu_{T_0} = n\mu$, $\sigma^2_{T_0} = n\sigma^2$. Quanto maior o valor de n, melhor a aproximação.

- De acordo com o TLC, quando n é grande e queremos calcular a probabilidade, tal como $P(a \leq \bar{X} \leq b)$, precisamos somente "fingir" que X é normal, padronizá-lo e usar a tabela normal.

- Regra prática: Se $n > 30$, o Teorema do Limite Central pode ser aplicado.

Teorema do Limite Central

Exemplo 1: Quando um lote de certo produto químico é preparado, a quantidade de uma impureza específica no lote é uma variável aleatória com valor médio de $4,0\ g$ e desvio padrão de $1,5\ g$. Se 50 lotes forem preparados independentemente, qual é a probabilidade (aproximada) de a quantidade média de impureza $\bar{X}$ da amostra estar entre $3,5$ e $3,8\ g$?

Teorema do Limite Central

Exemplo 1: Quando um lote de certo produto químico é preparado, a quantidade de uma impureza específica no lote é uma variável aleatória com valor médio de $4,0\ g$ e desvio padrão de $1,5\ g$. Se 50 lotes forem preparados independentemente, qual é a probabilidade (aproximada) de a quantidade média de impureza $\bar{X}$ da amostra estar entre $3,5$ e $3,8\ g$?

Resposta: De acordo com a regra prática, $n = 50$ é grande o suficiente para que o TLC seja aplicável. Então, $\bar{X}$ possui aproximadamente uma distribuição normal com valor médio $\mu_{\bar{X}} = 4,0$ e $\sigma_{\bar{X}} = 1,5/\sqrt{50} = 0,2121$. Assim,

$$
\begin{aligned}
P(3,5 < \bar{X} < 3,8) \approx & P\left(\frac{3,5-4,0}{0,2121} \leq Z \leq \frac{3,8-4,0}{0,2121}\right) \\
= & \Phi(-0,94) - \Phi(-2,36) = 0,1645. \blacksquare
\end{aligned}
$$

Teorema do Limite Central

Exemplo 2: Uma determinada organização de consumidores normalmente reporta o número de defeitos graves de cada carro novo examinado. Suponha que o número de tais defeitos de certo modelo seja uma variável aleatória com valor médio $3,2$ e desvio padrão $2,4$. Dentre 100 carros seleccionados aleatoriamente desse modelo, qual é a probabilidade de o número médio da amostra de defeitos graves exceder 4?

Teorema do Limite Central

Exemplo 2: Uma determinada organização de consumidores normalmente reporta o número de defeitos graves de cada carro novo examinado. Suponha que o número de tais defeitos de certo modelo seja uma variável aleatória com valor médio $3,2$ e desvio padrão $2,4$. Dentre 100 carros seleccionados aleatoriamente desse modelo, qual é a probabilidade de o número médio da amostra de defeitos graves exceder 4?

Resposta: Seja X_i o número de defeitos graves do i-ésimo carro na amostra aleatória.

- Observe que X_i é uma v.a. discreta,
- mas que o TLC é aplicável seja para v.a. discreta ou contínua.

Assim, para o tamanho $n = 100$ implica que $\bar{X}$ tem a distribuição aproximadamente normal com $\mu_{\bar{X}} = 3,2$ e $\sigma_{\bar{X}} = 2,4/\sqrt{100} = 0,24$.

$$P(\bar{X} > 4) \approx P\left(Z > \frac{4 - 3,2}{0,24}\right) = 1 - \Phi(3,33) = 0,0004. \blacksquare$$