Estatística aplicada

Lino Costa

Departamento de Produção e Sistemas
Escola de Engenharia
lac@dps.uminho.pt

Ano letivo 2018/2019

Sumário

População e amostra

População (ou Universo)

conjunto de indivíduos ou objetos que apresentam uma ou mais características em comum que se pretende analisar

Amostra

subconjunto da população, que se observa com o objetivo de se obter informação sobre características desconhecidas da população de onde foi retirada

Inferência estatística

- a população define um conjunto vasto, em geral, impossível de conhecer e a amostra constitui um subconjunto da população;
- uma amostra aleatória é uma amostra com n elementos independentes $x_1, x_2, \ldots, x_n$ em que a probabilidade de cada elemento ser selecionado é conhecida;
- o objetivo é, a partir da amostra aleatória, estabelecer conclusões para o todo representado pela população.

Parâmetro e estatística

Parâmetro

o parâmetro θ é uma característica numérica, em geral, constante que descreve a população

Exemplo 1

São exemplos de parâmetros a média (μ), a variância (σ^2), a proporção (p), o coeficiente de correlação (ρ)...

Estatística

a estatística é uma característica numérica, calculada a partir dos valores observados na amostra, que descreve a amostra

Exemplo 2

São exemplos de estatísticas a média amostral ($\bar{x}$), a variância amostral (s^2), a proporção amostral ($\hat{p}$), o coeficiente de correlação (R)...

Inferência estatística

A estatística inferencial refere-se a tomar decisões ou conclusões acerca da população através da análise da amostra

- estimação de parâmetros: estimar o valor de θ desconhecido
  - estimação pontual: estimar o valor exato de θ (por exemplo, $\mu = 10$)
  - estimação intervalar: estimar uma intervalo que inclua o verdadeiro valor de θ com uma determinada probabilidade (por exemplo, $7 < \mu < 13$)
- testes de hipóteses: testar uma hipótese acerca de θ (por exemplo, $H_0 : \mu = 10$)

Estimação pontual

Estimador pontual

A estimação pontual $\hat{\Theta}$ é uma estatística usada para estima θ. $\hat{\Theta}$ é uma variável aleatória porque uma estatística é uma variável aleatória

Exemplo 3

Estimador pontual de μ: média amostral $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Estimativa pontual

O valor numérico de $\hat{\Theta}$ calculado com base numa determinada amostra aleatória é chamada de estimativa pontual de θ (representada por $\hat{\theta}$).

Tendência de um estimador

Estimador não tendencioso

Um estimador não tendencioso (ou não enviesado ou centrado) é um estimador pontual $\hat{\Theta}$ cujo valor esperado é igual ao valor verdadeiro de θ.

$$E(\hat{\Theta}) = \theta$$

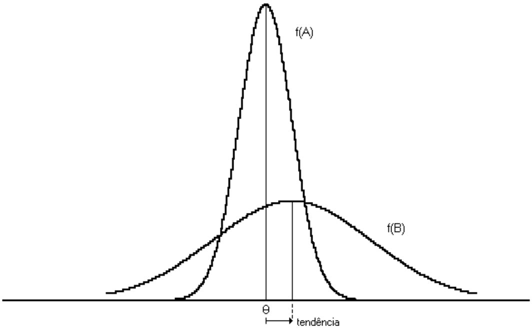
Tendência

A tendência de $\hat{\Theta}$ é a diferença entre o valor de esperado $E(\hat{\Theta})$ e o verdadeiro valor de θ.

$$t_{\hat{\Theta}}(\theta) = E(\hat{\Theta}) - \theta$$

Quanto menor a tendência de um estimador mais **exato** é o estimador.

Tendência de um estimador



Exemplo 4

- A é um estimador não tendencioso de θ, i.e., $t_A(\theta) = E(A) - \theta = 0$
- B é um estimador tendencioso de θ, i.e., $t_B(\theta) = E(B) - \theta \neq 0$

Tendência de um estimador

Exemplo 5

Mostre que $\frac{X}{n}$, sendo X o número de sucessos em n tentativas, é um estimador não tendencioso do parâmetro p da distribuição binomial.

- para uma distribuição binomial com parâmetros n e p, $\mu = E(X) = np$
- $E(\frac{X}{n}) = \frac{1}{n}E(X) = p$, logo $t_{\frac{X}{n}}(p) = p - p = 0$

Exemplo 6

Considere uma população com distribuição dada pela seguinte função densidade de probabilidade de uma variável aleatória x:
$f(x) = e^{-(x-\delta)}, \quad x > \delta$. Mostre que $\bar{x}$, a média amostral de uma amostra retirada da população é um estimador tendencioso de δ.

- $\mu = E(X) = \int_{\delta}^{\infty} xe^{-(x-\delta)}dx = [-xe^{-(x-\delta)}]_{\delta}^{\infty} - \int_{\delta}^{\infty} -e^{-(x-\delta)}dx = 1 + \delta$
- $E(\bar{X}) = \frac{E(X_1)+E(X_2)+\ldots+E(X_n)}{n} = \frac{nE(X)}{n} = E(X) = 1 + \delta \neq \delta$, logo $\bar{x}$ é um estimador tendencioso de δ
- $E(\bar{X} - 1) = \delta$, logo $\bar{x} - 1$ é um estimador não tendencioso δ

Variância de estimadores

Variância de um estimador

Quanto menor a variância $V(\hat{\Theta})$ de um estimador mais **preciso** é o estimador.

Estimador não tendencioso de variância mínima

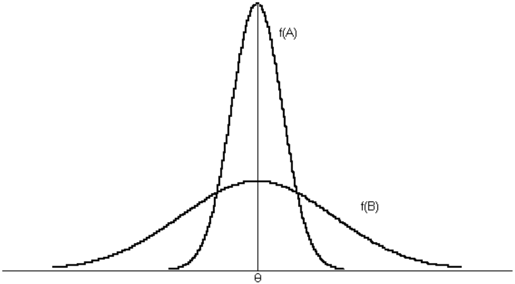
O estimador não tendencioso de variância mínima para θ é não tendencioso e de menor variância. É o estimador, simultaneamente, mais exato e preciso do parâmetro θ.

Eficiência relativa de dois estimadores

Se $\hat{\Theta}_1$ e $\hat{\Theta}_2$ são dois estimadores não tendenciosos do parâmetro θ de uma dada população e se a variância de $V(\hat{\Theta}_1)$ é menor que a variância de $V(\hat{\Theta}_2)$, diz-se que $\hat{\Theta}_1$ é relativamente mais eficiente que $\hat{\Theta}_2$.

$$efic(\hat{\Theta}_1, \hat{\Theta}_2) = \frac{V(\hat{\Theta}_1)}{V(\hat{\Theta}_2)}$$

Variância de estimadores



Exemplo 7

- A e B são estimadores não tendencioso de θ, i.e., $t_A(\theta) = E(A) - \theta = 0$ e $t_B(\theta) = E(B) - \theta = 0$, logo A e B são igualmente exatos.
- $V(A) < V(B)$, logo A é mais preciso do que B para estimar θ.
- $efic(A, B) = \frac{V(A)}{V(B)} < 1$, logo A é mais eficiente do que B.

Erro quadrático médio

Erro padrão de um estimador

O erro padrão de um estimador $\hat{\Theta}$ é o desvio padrão do estimador $\sigma_{\hat{\Theta}} = \sqrt{V(\hat{\Theta})}$. O erro padrão pode ser estimado por $s_{\hat{\Theta}}$, i.e., $\sigma_{\hat{\Theta}} \approx s_{\hat{\Theta}}$.

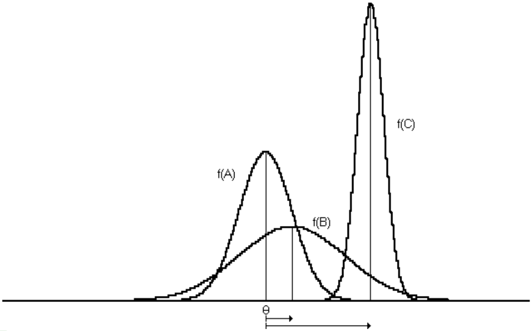
Erro Quadrático Médio (EQM)

Se $\hat{\Theta}$ não é um estimador não tendencioso de um dado parâmetro θ, as comparações devem ser feitas com base no erro quadrático médio ($EQM(\hat{\Theta})$) em vez de apenas a variância $V(\hat{\Theta})$.

$$EQM(\hat{\Theta}) = E\left((\hat{\Theta} - \theta)^2\right) = V(\hat{\Theta}) + \left[E(\hat{\Theta}) - \theta\right]^2$$

Para estimadores não tendenciosos de θ, $EQM(\hat{\Theta}) = V(\hat{\Theta})$.

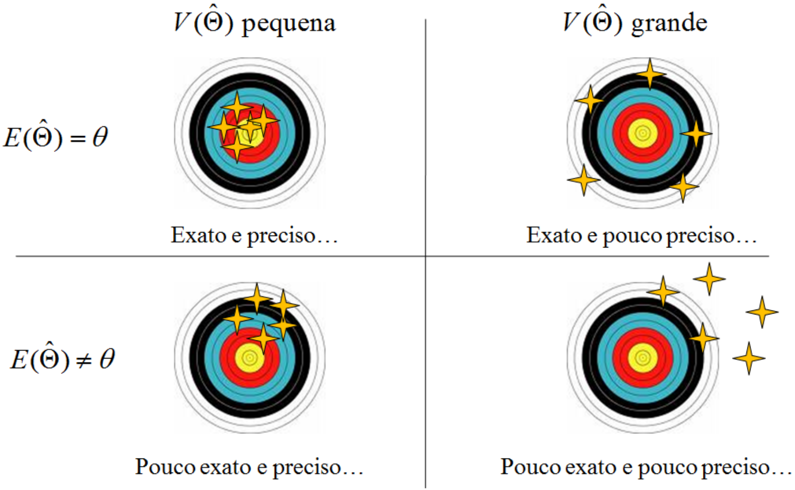
Erro quadrático médio



Exemplo 7

- A é um estimador não tendencioso de θ, B e C são estimadores tendenciosos de θ.
- C é mais preciso do que A e este mais preciso que B, i.e., $V(C) < V(A) < V(B)$.
- estimadores devem ser comparados em termos de EQM.

Erro quadrático médio



Consistência de um estimador

O estimador $\hat{\Theta}$ é um estimador consistente do parâmetro θ se e só se para cada $c > 0$

$$\lim_{n \to \infty} P\left(\left|\hat{\Theta} - \theta\right| < c\right) = 1$$

onde n é a dimensão da amostra aleatória. A consistência é uma propriedade assimtótica.
Se $\hat{\Theta}$ é um estimador não tendencioso do parâmetro θ e $V(\hat{\Theta}) \to 0$ à medida que $n \to \infty$, então $\hat{\Theta}$ é um estimador consistente de θ.

Exemplo 8

Mostre que a média amostral $\bar{x}$ é um estimador consistente da média da população μ.

- $V(\bar{X}) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{V(X_1)+V(X_2)+\ldots+V(X_n)}{n^2} = \frac{nV(X)}{n^2} = \frac{V(X)}{n} = \frac{\sigma^2}{n} \to 0$ quando $n \to \infty$, logo $\bar{x}$ é um estimador consistente de μ.

Suficiência de um estimador

O estimador $\hat{\Theta}$ é suficiente se usa toda a informação da amostra relevante para a estimação do parâmetro θ; i.e., se todo o conhecimento acerca de θ que pode ser ganho a partir dos valores amostrais individuais e da sua ordem, pode também ser ganho pelo valor de $\hat{\Theta}$ por si só.

Exemplo 9

O estimador $\hat{y} = \frac{1}{2n} \sum_{i=1}^{n} Y_i$ calculado para uma amostra aleatória de tamanho $2n$ não é suficiente pois não utiliza todo o conhecimento existente na amostra aleatória.

Consistência e suficiência

Consistência

$n \to \infty$

Suficiência

Suficiente…

Não suficiente…

Distribuição amostral

A distribuição amostral é a distribuição de probabilidade de uma estatística (uma função de variáveis aleatórias como a média amostral e a variância amostral. A distribuição amostral depende:

- da distribuição da população
- da dimensão da amostra
- do método de seleção da amostra

Exemplo 13

A distribuição amostral da média amostral $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$ para uma amostra aleatória de n observações independentes, $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$.

Distribuição amostral de $\bar{X}$

Considere uma amostra aleatória de dimensão n retirada de uma população normal com média μ e variância σ^2, então a distribuição amostral $\bar{X}$ é

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$$

Uma vez que $X_1, X_2, \ldots, X_n$ são independentes e normalmente distribuídas com a mesma média $E(X) = \mu$ e variância $V(X) = \sigma^2$, a distribuição amostral de $\bar{X}$ é normal com média e variância dadas por

$$\begin{aligned}
E(\bar{X}) &= E(\frac{X_1 + X_2 + \ldots + X_n}{n}) = \frac{1}{n}[E(X_1) + E(X_2) + \ldots + E(X_n)] = \\
&= \frac{1}{n}[\mu + \mu + \ldots + \mu] = \frac{n\mu}{n} = \mu \\
V(\bar{X}) &= V(\frac{X_1 + X_2 + \ldots + X_n}{n}) = \frac{1}{n^2}[V(X_1) + V(X_2) + \ldots + V(X_n)] = \\
&= \frac{1}{n^2}[\sigma^2 + \sigma^2 + \ldots + \sigma^2] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}
\end{aligned}$$

Teorema do limite central

Considere uma amostra aleatória $X_1, X_2, \ldots, X_n$ de n observações independentes retirada de uma qualquer população com média μ e variância σ^2, e então a distribuição limite da média amostral $\bar{X}$ é

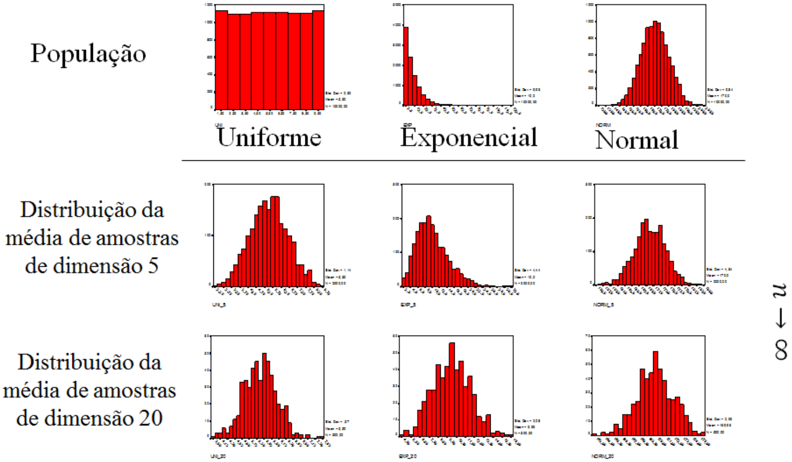
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

à medida que $n \to \infty$. Esta aproximação normal da distribuição de $\bar{X}$ é conhecida pelo Teorema do Limite Central.

Aplicação

- **População normal ($X \sim N(\mu, \sigma^2)$)** logo a distribuição de $\bar{X}$ é normal, i.e., $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- **População não normal com média μ e variância σ^2**
  - **Amostra grande ($n \geq 30$)** logo a distribuição de $\bar{X}$ é normal, i.e., $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
  - **Amostra pequena ($n < 30$)** logo a distribuição de $\bar{X}$ não é normal

Teorema do limite central



Teorema do limite central

Exemplo 14

Suponha que as classificações, a nível nacional, do exame de Geografia, têm uma média de 14.3, com um desvio padrão 2.1. Assumindo que a distribuição é normal, calcule a probabilidade de que:
i) um estudante, selecionado aleatoriamente, tenha uma classificação superior a 16 valores.
ii) uma amostra aleatória de 10 estudantes tenha uma média superior a 16 valores.

- A população é normal
- i) $X \sim N(\mu, \sigma^2)$ com $\mu = 14.3$ e $\sigma^2 = 2.1^2$
- $P(X > 16) = P\left(Z > \frac{16-14.3}{2.1}\right) = P(Z > 0.81) = 1 - P(Z \leq 0.81) = 1 - 0.7910 = 0.2090$
- ii) $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ com $\mu = 14.3$, $\sigma^2 = 2.1^2$ e $n = 10$
- $P(\bar{X} > 16) = P\left(Z > \frac{16-14.3}{\frac{2.1}{\sqrt{10}}}\right) = P(Z > 2.56) = 1 - P(Z \leq 2.56) = 1 - 0.9948 = 0.0052$

Teorema do limite central

Exemplo 15

Uma máquina de enchimento de açúcar está regulada por forma a que a quantidade em cada pacote seja de 1000 gramas, com um desvio padrão de 50 gramas.
Qual a probabilidade de que a média de uma amostra de 36 pacotes seja menor que 980 gramas?

- A população não é normal
- $n = 36 \geq 30 \Rightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ com $\mu = 1000$, $\sigma^2 = 50^2$ e $n = 36$
- $P(\bar{X} \leq 980) = P(Z \leq \frac{980-1000}{\frac{50}{\sqrt{36}}}) = P(Z \leq -2.41) = 0.0080$