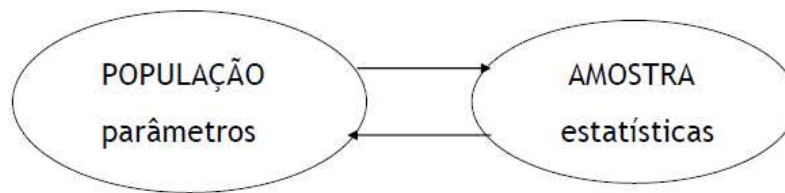


BIOESTATÍSTICA

Parte 4 – Estimação Pontual e Intervalar

Aulas Teóricas de 28/04/2011 a 05/05/2011

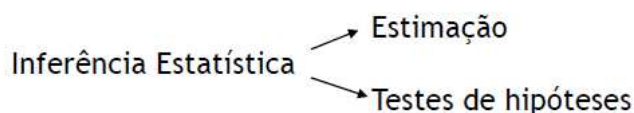
Inferência Estatística Paramétrica



Dada uma amostra, proveniente de uma população, o cálculo das características amostrais fornece-nos informação sobre as correspondentes características populacionais, ou seja, sobre os parâmetros (desconhecidos) da população.

[estatísticas (\bar{x} , s^2 , ...) caracterizam a amostra e os parâmetros (μ , σ^2) caracterizam a população e são muitas vezes desconhecidos. As últimas representam-se por letras gregas]

Utilizamos então as características amostrais (estatísticas) para inferir algo sobre a população de onde foi recolhida a amostra; estamos na fase da análise estatística designada por Inferência Estatística.



Inferir = tirar conclusões

4.1. Estimação Pontual: propriedades dos estimadores; método dos momentos; método da máxima verosimilhança

Quando um parâmetro é desconhecido, como podemos estimar ou “adivinhar” o seu verdadeiro valor, a partir da informação contida na amostra?

Estimação do Valor Médio

Consideremos uma população X cujo valor médio é desconhecido. Quando recolhemos uma amostra e obtemos a média, dizemos que este valor é uma estimativa do valor médio da v.a. X .

A v.a. \bar{X} é uma função da amostra aleatória que fornece estimativas para o valor médio; diz-se que é um estimador do valor médio.

\bar{X} é um estimador
 \bar{x} é uma estimativa } do valor médio

A média é um bom estimador do valor médio. Porquê?

Dada uma amostra aleatória (X_1, \dots, X_n) proveniente de uma população X de valor médio μ e variância σ^2 ,

$$E(\bar{X}) = \mu \quad e \quad \text{var}(\bar{X}) = \frac{\sigma^2}{n}$$

Então o valor médio do estimador coincide com o parâmetro que estamos a estimar. Diz-se que o estimador é centrado ou não enviesado.

Não é sistematicamente superior ou inferior ao valor verdadeiro

Isto significa que fornece estimativas que, em média, coincidem com o verdadeiro valor do parâmetro. Não haverá tendência sistemática para sobrestimar ou subestimar o valor do parâmetro.

Além disso, a variância do estimador decresce com a dimensão da amostra, portanto, à medida que aumentamos a dimensão da amostra, a variabilidade do estimador, em torno do parâmetro, diminui. Diz-se então que o estimador é consistente.

$\frac{\sigma}{\sqrt{n}}$ designa-se por erro padrão da média

Estimação da Variância Populacional

A justificação para a utilização da variância amostral dada por

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

como estimador da variância populacional é: $E(S^2) = \sigma^2$

S^2 é um estimador centrado da variância populacional

Estimação da Proporção

Consideremos uma população constituída por indivíduos que pertencem a uma de duas categorias A e \bar{A} . Seja p a proporção (desconhecida) de indivíduos que pertencem à categoria A.

Pretendemos fazer inferência sobre o parâmetro p , pelo que se recolhe da população uma amostra de dimensão n . Seja X a v.a. que representa o número de indivíduos na amostra que pertencem à categoria A.

Um estimador natural para p é a frequência relativa X/n , ou seja, a proporção de indivíduos da amostra que pertencem à categoria A.

$$\hat{p} = \frac{X}{n} \quad \text{estimador} \qquad \hat{p} = \frac{x}{n} \quad \text{estimativa}$$

Distribuição de amostragem da proporção \hat{p}

$$\hat{p} = \frac{X}{n} \text{ e } X \sim \text{Bi}(n, p) \quad \text{então} \quad E(\hat{p}) = p \quad \text{var}(\hat{p}) = \frac{p(1-p)}{n}$$

Pelo Teorema Limite Central, para n suficientemente grande, a distribuição de amostragem de \hat{p} pode ser aproximada pela distribuição normal, de valor médio p e variância $p(1-p)/n$

4.2. Estimação Intervalar: intervalos de confiança para o valor médio e variância de uma população gaussianana; comparação dos parâmetros de duas gaussianas; intervalos de confiança assintóticos em população não gaussianas; intervalos de confiança para uma proporção e para a diferença entre proporções

Em vez de propormos um único valor como estimativa de um parâmetro, podemos obter, com base na amostra, um intervalo de valores ao qual atribuímos um certo **grau ou coeficiente de confiança** em como contém o verdadeiro valor (desconhecido) desse parâmetro.

Diz-se então que vamos obter uma estimativa intervalar, ou seja, um intervalo de confiança para o parâmetro.

Intervalo de Confiança para o Valor Médio

Seja $X \sim N(\mu, \sigma)$ com μ desconhecido e σ conhecido e consideremos uma amostra aleatória (X_1, \dots, X_n) proveniente da população X .

$$\text{Então} \quad Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Z é uma v.a. cuja distribuição não depende do parâmetro desconhecido μ .
Designa-se Z por **variável fulcral**.

Fixado o grau, coeficiente ou nível de confiança $1-\alpha$ ($0 < \alpha < 1$), tem-se que

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

Temos assim o intervalo aleatório

$$\left(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

Para uma amostra observada (x_1, \dots, x_n) , obtém-se então um intervalo de confiança a $100(1-\alpha)\%$ ou intervalo de $100(1-\alpha)\%$ de confiança para o valor médio μ

$$\left(\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

O intervalo de confiança vem na forma $(\bar{x} - \varepsilon, \bar{x} + \varepsilon)$, em que ε é a margem de erro, que é metade da amplitude do intervalo

Ex: O tempo (em minutos) que uma determinada tarefa leva a executar segue uma distribuição normal com desvio padrão 2. Foi observada uma amostra de dimensão 16, em que a média obtida foi 12.5 minutos. Construa um intervalo de 95% de confiança para o valor médio μ .

$$1-\alpha = 0.95 \Leftrightarrow \alpha = 0.05 \Leftrightarrow 1-\alpha/2 = 0.975$$

$$z_{0.975} = 1.96$$

Um intervalo de 95% de confiança para μ (com σ conhecido) é da forma

$$\left(\bar{x} - z_{0.975} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.975} \frac{\sigma}{\sqrt{n}}\right)$$

e neste caso obtemos o seguinte intervalo

$$\left(12.5 - 1.96 \times \frac{2}{\sqrt{16}}, 12.5 + 1.96 \times \frac{2}{\sqrt{16}}\right) = (11.52, 13.48)$$

Com 95% de confiança, o tempo médio de execução da tarefa situa-se entre 11.52 e 13.48 minutos. Estimámos o valor médio com uma margem de erro igual a 0.98.

O valor médio poderá ser igual a 10? Com 95% de confiança não, porque 10 não pertence ao intervalo. Note-se que é errado dizer pragmaticamente que o valor 10 não é o verdadeiro valor médio, pois este continua desconhecido, mesmo após a inferência.

Interpretação do Intervalo de Confiança

Ao calcular um intervalo de $100(1-\alpha)\%$ de confiança para μ com base numa amostra que foi recolhida da população, o intervalo obtido contém ou não contém o verdadeiro valor do parâmetro.

Se forem recolhidas várias amostras da mesma dimensão, o valor da média será certamente diferente de amostra para amostra, pelo que os limites do intervalo variam. Portanto, a cada uma das amostras observadas corresponde um intervalo particular dado por

$$\left(\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

Então, se for observado um grande número de amostras de dimensão n , a percentagem destes intervalos que contém μ é aproximadamente $100(1-\alpha)\%$. Portanto, **esperamos que $100(1-\alpha)\%$ dos intervalos obtidos contenham μ , enquanto que os restantes $100\alpha\%$ não o conterão.**

Ex: Para um nível de confiança de 95%, se forem recolhidas 100 amostras da mesma dimensão de uma determinada população e se calcularmos os intervalos correspondentes, esperamos que 95 desses intervalos contenham o verdadeiro valor do parâmetro.

- Quanto maior a amplitude do intervalo, menos a precisão e maior a margem de erro;
- Para um maior nível de confiança, mantendo a amplitude do intervalo, só recolhendo uma amostra maior.

Para uma determinada amostra, quanto maior for o nível de confiança exigido, maior será a amplitude do intervalo obtido.

Um processo para diminuir a amplitude do intervalo de confiança (i.e., aumentar a precisão da estimativa) consiste em aumentar a dimensão da amostra.

Seja $X \sim N(\mu, \sigma)$ com μ desconhecido e σ conhecido. Qual a dimensão da amostra necessária para obter um intervalo de $100(1-\alpha)\%$ de confiança para o valor médio cuja amplitude não exceda um certo valor d ?

$$2z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq d \Leftrightarrow n \geq \left(\frac{2z_{1-\alpha/2} \sigma}{d} \right)^2$$

Se σ for desconhecido, para determinar a dimensão da amostra que vai ser recolhida, substitui-se σ pelo valor de s obtido através de um estudo piloto.

Designa-se por **margem de erro** o valor igual a metade da amplitude do intervalo de confiança.

Intervalos de $100(1-\alpha)\%$ de confiança para o valor médio

População normal

σ conhecido $\left(\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$

σ desconhecido $\left(\bar{x} - t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}} \right)$

$t_{n-1, 1-\alpha/2}$ é o quantil de probabilidade $1-\alpha/2$ da distribuição t-Student com $n-1$ graus de liberdade

População não normal, grande amostra ($n \geq 30$)

σ conhecido $\left(\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$

σ desconhecido $\left(\bar{x} - z_{1-\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \frac{s}{\sqrt{n}} \right)$

Intervalo de Confiança para a Variância da População

População normal, μ desconhecido

A variável fulcral é $\frac{(n-1)S^2}{\sigma^2} \cap \chi^2_{n-1}$

Intervalo de $100(1-\alpha)\%$ de confiança para a variância σ^2

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1, 1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1, \alpha/2}} \right)$$

$\chi^2_{n-1, \alpha}$ é o quantil de probabilidade α da distribuição Qui-quadrado com $n-1$ graus de liberdade

Intervalo de Confiança para a proporção p (Grandes Amostras)

Fixado o nível de confiança $1-\alpha$, recordando a distribuição de amostragem de \hat{p} para n suficientemente grande, tem-se que

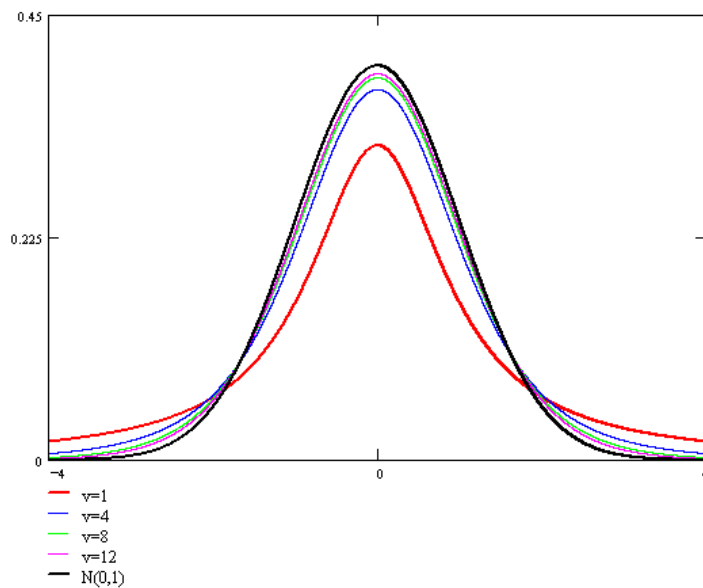
$$P\left(-z_{1-\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq z_{1-\alpha/2}\right) \approx 1-\alpha$$

Substituindo p , nos extremos do intervalo, pelo seu estimador consistente \hat{p} , obtém-se o seguinte intervalo (aproximado) de $100(1-\alpha)\%$ de confiança para p

$$\left(\hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

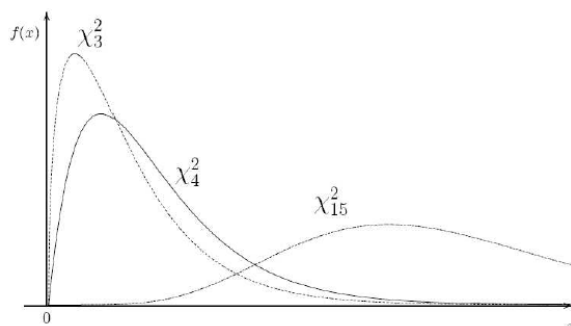
Nota:

Distribuição t-Student



Há medida que $n-1$ graus de liberdade aumenta, f.d.p. aproxima-se da $N(0,1)$

Distribuição Qui-Quadrado (χ^2)



- Não é uma distribuição simétrica;
- É sempre positiva.