UAlg esght

Distribuições por amostragem

Paulo Batista Basílio
(pbasilio@ualg.pt)

Dezembro 2014

DISTRIBUIÇÃO AMOSTRAL DA MÉDIA

Um processo de amostragem conduz virtualmente a muitas amostras diferentes:

$$Populao \Rightarrow \begin{cases} Amostra\ 1 \rightarrow \hat{\theta} = T(x_{11}, x_{21}, \ldots, x_{n1}) \\ Amostra\ 2 \rightarrow \hat{\theta} = T(x_{12}, x_{22}, \ldots, x_{n2}) \\ \qquad \vdots \\ Amostra\ m \rightarrow \hat{\theta} = T(x_{1m}, x_{2m}, \ldots, x_{nm}) \end{cases}$$

Neste contexto, uma estatística é uma variável aleatória, $T(X_1, X_2, ?, X_n)$ que não envolve qualquer parâmetro desconhecido. Portanto, a representação apropriada será

$$\hat{\Theta} = T(X_1, X_2, \ldots, X_n)$$

onde $X_1, X_2, ..., X_n$ são variáveis aleatórias, representando a amostra da variável X (população).

É através da distribuição de amostragem que é introduzida a probabilidade num procedimento estatístico e podemos inferir conclusões para uma população a partir de resultados amostrais.
A distribuição da amostra traduz a estrutura da população de amostras de dimensão n obtidas da população representada pela variável aleatória X e é dada pela função densidade (função probabilidade) conjunta

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

Consideremos o seguinte exemplo. Uma urna contém três bolas numeradas com os números 2, 4 e 6. São extraídas, com reposição, amostras de dimensão n = 2, a distribuição amostral da média $(\bar{X})$ pode ser calculada da seguinte forma:
1. A distribuição de probabilidade do universo é uma distribuição discreta e uniforme, onde a probabilidade de ser escolhida uma bola ao acaso é 1/3.
2. O número total de amostras possíveis é 9.

$Amostra$	$2,4$	$2,2$	$4,2$	$2,6$	$6,6$	$6,2$	$4,6$	$4,4$	$6,4$
$\bar{X}$	3	2	3	4	6	4	5	4	5

3. A distribuição de probabilidade será

$\bar{X}$	2	3	4	5	6
$P(\bar{X} = \bar{x})$	0.111	0.222	0.333	0.222	0.111

onde,

$$P(\bar{X}=2) = P\left(X_1 = 2 \cap X_2 = 2\right) = \frac{1}{3} \times \frac{1}{3} = 0.111$$

$$P(\bar{X}=3) = P\left[(X_1 = 2 \cap X_2 = 4) \cup (X_1 = 4 \cap X_2 = 2)\right] = \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3} = 0.222$$

$$P(\bar{X}=4) = P\left[(X_1 = 2 \cap X_2 = 6) \cup (X_1 = 6 \cap X_2 = 2) \cup (X_1 = 4 \cap X_2 = 4)\right] = \frac{1}{3}\times\frac{1}{3}+\frac{1}{3}\times\frac{1}{3}+\frac{1}{3}\times\frac{1}{3} = 0$$

$$P(\bar{X}=5) = P\left[(X_1 = 4 \cap X_2 = 6) \cup (X_1 = 6 \cap X_2 = 4)\right] = \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3} = 0.222$$

$$P(\bar{X}=6) = P\left(X_1 = 6 \cap X_2 = 6\right) = \frac{1}{3} \times \frac{1}{3} = 0.111$$

Neste exemplo, a distribuição amostral diz-nos que se extrairmos 2 bolas a média mais provável que obteremos é 4, porque corresponde a três amostras em nove possíveis.

Neste contexto, uma estatística (por exemplo, a média) é uma variável aleatória que depende da amostra seleccionada, que não inclui qualquer parâmetro desconhecido, e com uma determinada distribuição por amostragem dada pela respectiva função distribuição. Assim, ao conjunto de valores possíveis da estatística será possível ajustar uma função (densidade) de probabilidade com parâmetros geralmente desconhecidos. No entanto, para estatísticas como a média e a variância temos

Teorema 1. Se $(X_1, X_2, ?, X_n)$ é uma amostra casual de população para a qual existem média $\mu = E(X_i)$ e variância $\sigma^2 = Var(X_i)(i = 1, 2, ..., n)$, tem-se

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

Repare-se que no exemplo anterior

? A média do universo é $\mu = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{2+4+6}{3} = 4$
? A variância $\sigma^2 = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n} = \frac{(2-4)^2+(4-4)^2+(6-4)^2}{3} = 2.67$
? A média das médias das 9 amostras é igual a 4, ou seja, $\mu_{\bar{X}} = \mu$
? O desvio padrão das médias amostrais é igual a 1.155, ou seja,

$$\sigma_{\bar{X}} = \sqrt{\sum_{i=1}^{n}(X_i - \mu)^2 f(x_i)} = \sqrt{(2-4)^2 \times 0.111 + \ldots + (6-4)^2 \times 0.1111} = 1.155 = \frac{\sigma}{n}$$

Em universos e amostras tão limitados é relativamente fácil estudar todas as possibilidades, mas quando os universos permitem amostras maiores a questão já não será tão

simples. Em situações em que a amostra é relativamente grande entra em cena um dos teoremas mais importantes da estatística que suporta toda a inferência clássica. O teorema do limite central que relaciona o aumento da dimensão da amostra com a normalidade da distribuição da média amostral.

Teorema do limite central. Dada a sucessão de variáveis aleatórias iid, $X_1, X_2, \ldots, X_n$ com média μ e variância σ^2 então, quando $n \to \infty$ a função de distribuição da variável aleatória,

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$$

tende para uma função de distribuição N(0,1).

Corolário. Dada a sucessão de variáveis aleatórias iid, $X_1, X_2, \ldots, X_n$ com média μ e variância σ^2, então

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