Distribuições Amostrais

Prof. Eduardo Bezerra

Inferência Estatística

31 de agosto de 2018

Roteiro

Roteiro

Distribuições amostrais

Considere a realização de uma amostragem aleatória simplesde uma população para produzir uma amostra de n elementos.

- Uma afirmação eventualmente feita sobre essa população será baseada em alguma estatística T, que é uma função da amostra $(X_1, X_2, \ldots, X_n)$.
- Colhida essa amostra, teremos observado um valor particular de T.

A validade dessa afirmação seria melhor compreendida se soubéssemos o que acontece quando produzimos **todas** as amostras da população.

- Isto é, qual é a distribuição de T quando $(X_1, X_2, \ldots, X_n)$ assume todos os valores possíveis.

Distribuições amostrais (cont.)

A distribuição obtida considerando todas as possíveis amostras de uma população é denominada distribuição amostral (*sampling distribution*) da estatística T.

Distribuições amostrais (cont.)

O procedimento geral para obtenção dessa distribuição envolve os seguintes componentes:

(a) uma população X, com determinado parâmetro de interesse θ;
(b) todas as amostras retiradas da população, de acordo com certo plano amostral;

Para cada amostra, calculamos o valor t (*estatística pontual*) da estatística T.

Os valores t formam uma nova população, cuja distribuição recebe o nome de distribuição amostral de T.

Roteiro

Distribuição amostral da média

Quando a estatística sendo calculada é a média amostral (*sample mean*), a distribuição amostral é denominada distribuição amostral da média (*sample mean distribution*).

Distribuição amostral da média

Suponha que uma amostra aleatória de tamanho n seja retirada de uma população normal com média μ e variância σ^2.
Pela *propriedade reprodutiva* da distribuição normal, a estatística denominada média amostral

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

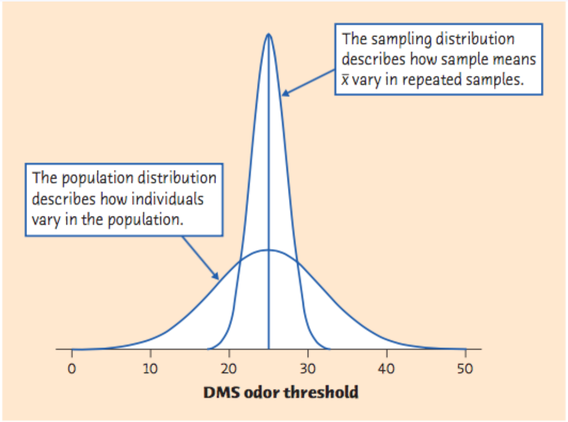
tem uma distribuição normal com média

$$\mu_{\overline{X}} = \frac{1}{n}(\mu + \mu + \ldots + \mu) = \mu$$

e variância

$$\sigma^2_{\overline{X}} = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \ldots + \sigma^2) = \frac{\sigma^2}{n}.$$

Distribuição amostral da média



Roteiro

Distribuição amostral das proporções

Considere uma população em que a proporção de elementos portadores de certa característica é p. Logo, podemos definir uma v.a. X da seguinte maneira:

$$X = \begin{cases} 1 & \text{se o elemento for portador da característica} \\ 0 & \text{se o elemento for não portador da característica} \end{cases}$$

A v.a. X assim definida segue a distribuição de Bernoulli, i.e.,

$$\mu = E(X) = p,\ \sigma^2 = Var(X) = p(1-p)$$

Distribuição amostral da proporção (cont.)

Se retirarmos uma AAS dessa população, e indicarmos por Y_n o total de elementos portadores da característica estudada **na amostra**, então

$$Y_n = X_1 + X_2 + \ldots + X_n$$

onde:

- cada X_i tem distribuição de Bernoulli;
- X_i e X_j, $i \neq j$, são independentes.

Por outro lado, sabemos que Y_n assim definida segue uma distribuição binomial, i.e.,

$$Y_n \sim b(n,p)$$

em que p é a probabilidade de um elemento ser portador da característica.

Distribuição amostral das proporções (cont.)

Vamos definir a v.a. $\widehat{p}$ como sendo a proporção de elementos na amostra que são portadores da característica, i.e.,

$$\widehat{p} = \frac{Y_n}{n}$$

Então

$$\Pr(Y_n = k) = \Pr(Y_n/n = k/n) = \Pr(\widehat{p} = k/n)$$

ou seja, a distribuição amostral de $\widehat{p}$ é obtida da distribuição de Y_n.

Distribuição amostral da proporção (cont.)

Já que $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$, podemos escrever

$$Y_n = n\overline{X}$$

Mas, pelo TLC, $\overline{X}$ possui distribuição aproximadamente normal, com média p e variância $(p(1-p))/n$, ou seja,

$$\overline{X} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Distribuição amostral da proporção (cont.)

Observe que, na expressão anterior, a variável $\overline{X}$ é a própria variável $\widehat{p}$ e, desse modo, para n grande, o TLC nos permite considerar a distribuição amostral de $\widehat{p}$ como sendo aproximadamente normal:

$$\widehat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Conclusão: a proporção amostral, assim como a média amostral, segue a distribuição normal.

Roteiro

Determinação do tamanho da amostra

- Até aqui, consideramos que o tamanho da amostra, n, é conhecido.
- Entretanto, podemos em certas ocasiões querer determinar o *tamanho mínimo da amostra* a ser colhida de uma população, de modo a obter um erro de estimação previamente estipulado, com determinado grau de confiança.

Determinação do tamanho da amostra (cont.)

Por exemplo, considere que estejamos estimando a média populacional μ e para tanto usaremos a média amostral, $\overline{X}$, baseada em uma amostra de tamanho n. Suponha que se queira determinar o valor de n de modo que

$$\Pr(|\overline{X} - \mu| \leq \epsilon) \geq \gamma$$

onde:

- γ é o grau de confiança $(0 < \gamma < 1)$;
- ϵ é o erro amostral máximo (ou erro de estimação) que podemos suportar.

Ambos os valores, γ e ϵ, são fixados durante o planejamento da pesquisa estatística.

Determinação do tamanho da amostra (cont.)

Sabemos que $\overline{X} \sim N(\mu, \sigma^2/n)$, logo $\overline{X} - \mu \sim N(0, \sigma^2/n)$ e portanto podemos reescrever a equação anterior como

$$\Pr(-\epsilon \leq \overline{X} - \mu \leq \epsilon) = \Pr\left(\frac{-\sqrt{n}\epsilon}{\sigma} \leq Z \leq \frac{\sqrt{n}\epsilon}{\sigma}\right) \geq \gamma$$

com $Z = (\overline{X} - \mu)\sqrt{n}/\sigma$. Dado γ, podemos obter z_γ da $N(0{,}1)$, tal que $\Pr(-z_\gamma < Z < z_\gamma) = \gamma$, de modo que

$$\frac{\sqrt{n}\epsilon}{\sigma} = z_\gamma$$

do que obtemos

$$n = \frac{\sigma^2 z_\gamma^2}{\epsilon^2}$$

Determinação do tamanho da amostra (cont.)

Portanto, o tamanho mínimo de uma amostra pode ser obtido por

$$n = \frac{\sigma^2 z_\gamma^2}{\epsilon^2}$$

Note que nessa expressão derivada para n, são conhecidos z_γ e ϵ, mas σ^2 é a variância populacional desconhecida. Portanto, para calcularmos n,

- devemos ter alguma informação prévia sobre σ^2 (obtida a partir de estudos anteriores ou pela experiência do projetista do estudo estatístico),
- ou então usar uma pequena amostra piloto para estimar σ^2.

Determinação do tamanho da amostra - exemplo 01

Suponha que uma pequena amostra piloto de $n = 10$, extraída de uma população, forneceu valores $\overline{X} = 15$ e $S^2 = 16$. Com valores fixos de $\epsilon = 0{,}5$ e $\gamma = 0{,}95$, temos

$$n = \frac{16 \times (1{,}96)^2}{(0{,}5)^2} = 245.$$

Na expressão acima, 1,96 é o valor na normal padrão necessário para que se obtenha 95% da área sob a curva.

Determinação do tamanho da amostra

Quando o parâmetro que se deseja estimar é p, a proporção populacional, podemos usar a aproximação normal para $\widehat{p}$, o que resulta em

$$n = \frac{z_\gamma^2 p(1-p)}{\epsilon^2}$$

Quando não conhecemos o valor de p, podemos usar o fato de que $p(1-p) \leq 1/4$, para todo p, de modo que

$$n \approx \frac{z_\gamma^2}{4\epsilon^2}$$

Determinação do tamanho da amostra - exemplo 02

Em uma pesquisa de mercado, estima-se que aproximadamente 60% das pessoas entrevistadas preferirão a marca A de um produto. Essa informação é baseada em pesquisas anteriores. Considere que, no planejamento da pesquisa, definiu-se que o erro amostral máximo de $\widehat{p}$ deve ser 0,03 (i.e., $\epsilon = 0{,}03$), com probabilidade $\gamma = 0{,}95$. Com os valores fornecidos para ϵ e γ, temos

$$n = \frac{(1{,}96)^2(0{,}6)(0{,}4)}{(0{,}03)^2} = 1.024.$$

em que usamos o fato de que $p \approx 0{,}60$.