4.5.3. Distribuições amostrais da média (amostras grandes)

Na seção anterior, fizemos por simulação uma estimativa empírica da *distribuição amostral* de $\bar{X}$; isto é, da distribuição das médias $\bar{X}$ das amostras que são retiradas de uma população de média conhecida. Veremos agora um modelo probabilístico para esta distribuição.

Suponha que desejamos testar uma hipótese sobre a média de um população desconhecida (representaremos esta média pela letra grega μ – pronuncia-se *mi* ou *mu*). Como vimos na seção anterior, se retiramos uma amostra aleatória da população e calculamos sua média ($\bar{x}_1$), não podemos simplesmente dizer que a média μ da população é igual à média da amostra, isto é, que $\mu = \bar{X}_1$. Se retirarmos outras amostras de mesmo tamanho, provavelmente veremos que todas terão médias diferentes entre si (e provavelmente todas diferentes de μ). Para chegarmos a alguma conclusão, precisamos compreender *como* variam estas médias das amostras; isto é, precisamos descobrir se há alguma lógica, algum padrão nesta variação. Para isto, iremos tratar estas médias $\bar{X}$ como observações de uma variável aleatória, e procurar um modelo para a distribuição de probabilidades desta variável (um dos modelos de VAC vistos na seção 3.4). A partir deste modelo, poderemos calcular as probabilidades associadas a cada intervalo de valores de $\bar{X}$.

Note que o problema está se complicando, e precisamos tomar cuidado com a notação. Estamos agora nos referindo a três distribuições diferentes:

- a distribuição da variável X na população;
- a distribuição da variável X na amostra;
- a distribuição teórica das médias $\bar{X}$ de todas as amostras possíveis.

Temos portanto três distribuições, com três médias e três desvios-padrões diferentes. No exemplo do peso dos peixes:

- na *população* de peixes, cada peixe tem um peso diferente; este peso é uma variável aleatória, que representaremos por X; o que queremos conhecer é a média μ desta variável (seu valor esperado);
- numa *amostra*, cada elemento (peixe) tem um peso diferente. A média dos pesos dos peixes numa amostra é representada por $\bar{X}$;
- cada amostra que retirarmos terá uma média diferente. A amostra 1 terá média $\bar{X}_1$; a amostra 2 terá média $\bar{X}_2 \neq \bar{X}_1$, e assim por diante. Portanto, $\bar{X}$ também será uma VAC. Para chegar a alguma conclusão, precisamos de encontrar um modelo para esta variável.

Para organizar isto tudo, precisamos de definir alguns conceitos. Chamamos de *estatística amostral* qualquer valor calculado como função dos dados da amostra (por exemplo, a média $\bar{X}$ da amostra ou sua proporção P de sucessos); chamamos de *parâmetros* os valores correspondentes na população (por exemplo, sua média μ ou sua proporção de sucessos π). Em geral, iremos representar os parâmetros por letras gregas (por exemplo, μ e π), e as estatísticas amostrais por letras romanas ($\bar{X}$ e P). É a partir destas *estatísticas* (com *e* minúsculo) que a *Estatística* (com *E* maiúsculo) vai fazer *inferência* sobre os parâmetros: fazer *testes de hipóteses* sobre eles (nas próximas seções), ou estimar seus valores por meio de *intervalos de confiança* (seção 4.8).

A distribuição de probabilidades da variável $\bar{X}$ é chamada de *distribuição das médias das amostras* (óbvio!) ou de *distribuição amostral das médias* (menos óbvio... Tradução do inglês *sampling distribution*). Usaremos esta segunda denominação, por ser a mais comum no Brasil. A distribuição de qualquer estatística amostral é chamada de *distribuição amostral* daquela estatística. Não só a média, mas qualquer estatística calculada em uma amostra (proporção de

sucessos, desvio-padrão, mediana, valores máximos e mínimo, etc.) tem a sua própria distribuição amostral. O Quadro **1** mostra como as três distribuições se relacionam, e os símbolos mais usuais, para o caso de um teste de média.

Quadro 1. Notação de estatísticas amostrais e parâmetros

	Símbolo		
	População	Amostra *i-ésima*	Distribuição amostral de $\overline{X}$
variável	X	X	$\overline{X}$
média	μ	$\overline{X}_i$	$\mu_{\overline{X}}$
desvio-padrão	σ	s_i	$\sigma_{\overline{X}}$

4.5.3.1. Teoremas sobre a distribuição amostral das médias

Se retiramos amostras de uma população normal (gaussiana), a média destas amostras também terá distribuição normal. Esta é uma das propriedades da distribuição normal (seção 3.4.4): a soma de variáveis independentes de distribuição normal é uma variável normal. A média nada mais é do que uma soma de variáveis (cada elemento da amostra é uma observação de uma variável normal), dividida por uma constante (o tamanho *n* da amostra); portanto, a média também é normal. Esta propriedade pode ser enunciada como o Teorema **1**.

Teorema 1. *Distribuição de das médias $\overline{X}$ de amostras de população normal*
Se amostras aleatórias simples de tamanho *n* (qualquer) são retiradas de uma população normal de média μ e desvio-padrão σ, a média amostral $\overline{X}$ será uma variável com distribuição normal,

$$\overline{X} \sim N(\mu_{\overline{X}}, \sigma^2_{\overline{X}})$$

cujos parâmetros são: $\mu_{\overline{X}} = \mu$ e $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$

Portanto, a variável padronizada $Z \sim N(0,1)$ será dada por

$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \rightarrow z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

Na prática, este teorema é bastante restritivo, pois nem sempre as populações que nos interessam têm distribuição normal. No entanto, há o Teorema **2** (provavelmente o teorema mais importante da Inferência Estatística, publicado por Laplace em 1810) que diz que se as amostras forem grandes a distribuição amostral de $\overline{X}$ tenderá para a distribuição normal, mesmo que a população não seja normal:

Teorema 2. *Teorema do Limite Central*
Se amostras aleatórias simples de tamanho *n* são retiradas de uma população (de qualquer distribuição) de média μ e desvio-padrão σ, a média amostral $\overline{X}$ será uma variável que tende para uma distribuição normal:

$$\overline{X} \rightarrow N(\mu_{\overline{X}}, \sigma^2_{\overline{X}}) \text{ quando n} \rightarrow \infty$$

na qual $\mu_{\overline{X}} = \mu \qquad \sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$

Portanto, a variável de teste padronizada Z :

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\overline{\sigma}} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

irá tender para a distribuição normal padrão $Z \sim N(0,1)$ quando for grande (n $\rightarrow\infty$). Este valor padronizado é chamado de *estatística de teste*, pois é a partir dele que tomaremos a decisão no teste.

Em geral, a variância populacional σ usada na fórmula acima é desconhecida; podemos contudo usar a variância *s* da amostra como uma estimativa da variância da população; a aproximação será razoável, desde que a amostra seja grande. Faremos então:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

onde

$$s = \sqrt{\frac{\sum (x - \overline{X})^2}{n}}$$

Note que o Teorema **2** não faz nenhuma exigência sobre a forma da população; apenas exige que a amostra seja aleatória simples e que *n* seja grande (na prática, alguns autores sugerem que uma amostra pode ser considerada suficientemente grande quando *n*>30; outros são mais conservadores, e preferem *n*>50). Note também que os dois teoremas acima chegam aos mesmos parâmetros da distribuição amostral ($\mu_{\overline{X}}$ e $\sigma_{\overline{X}}$), mas têm pressupostos diferentes: o primeiro exige que a população seja normal, o segundo exige que a amostra seja grande.

4.5.3.2. Exemplo de teste de média

Para exemplificar o uso do Teorema **2**, voltemos ao problema mostrado na seção 4.5.1, sobre o peso dos peixes de uma certa espécie; repetimos abaixo seu enunciado:

> (3) Devido ao excesso de pesca, surgiu a hipótese de que o peso médio dos peixes de uma espécie comum no Atlântico Norte está diminuindo, porque estes peixes não têm tempo de crescer suficientemente antes de serem pescados. Esta espécie tinha anteriormente peso médio de 28,0 kg, com desvio padrão de 4,0 kg. Um pesquisador retira uma amostra aleatória de 60 destes peixes, e encontra um peso médio de 26,0 kg. Este resultado é uma evidência de que o peso dos peixes está diminuindo?

Na seção 4.5.2 concluímos, por simulação em computador, que seria pouco provável que uma amostra aleatória de n=60 da população destes peixes tivesse peso igual ou menor que 26,0 kg. Podemos agora refazer este exemplo, usando o Teorema do Limite Central.

As hipóteses consideradas no teste serão:

$$H_0: \quad \mu \geq 28{,}0$$

$$H_1: \quad \mu < 28{,}0$$

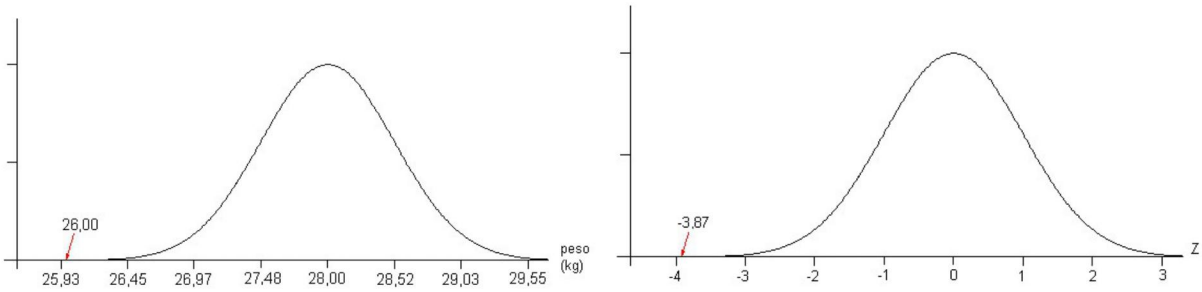
Se a população dos peixes tinha anteriormente peso médio μ=28,0 kg, com desvio-padrão σ = 4,0 kg, com distribuição não especificada (não necessariamente normal), e retiramos uma

amostra grande, o Teorema **2** diz que a média das amostras irá tender para uma VA normal de parâmetros

$$\mu_{\overline{X}} = \mu = 28{,}0$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4{,}0}{\sqrt{60}} = 0{,}5164$$

Esta distribuição amostral normal está representada na Fig. **1A**. Nela podemos ver que seria muito pouco provável encontrarmos uma amostra cuja média seja de $\overline{X} = 26{,}0$ kg.



A. Variável original $\overline{X}$ (média amostra, em kg) B. Variável padronizada Z

Figura 1. Distribuição amostral de $\overline{X}$ (variável: peso de peixes)

A média amostral de 26,0 kg corresponde a um valor padronizado de:

$$Z = \frac{26 - 28}{0{,}5164} = -3{,}87$$

A probabilidade de um valor igual ou menor que este ser encontrado é muito pequena:

$$P(Z \leq -3{,}87) = P(\overline{X} \leq 26) = 0{,}00005$$

Este valor de Z (representado na Fig. **1B**) confirma a conclusão obtida visualmente na Fig. **1A**.

Para usar o Teorema do Limite Central num problema real de Inferência, temos que fazer duas aproximações. Primeiro, o valor real da média populacional μ geralmente não é conhecido; nos *testes estatísticos*, esta média será dada por uma *hipótese*, que é o que queremos testar. Além disso, o desvio-padrão populacional σ também em geral não é conhecido, e teremos que estimar seu valor a partir do desvio-padrão s da amostra.

Se a amostra for grande, a estimação pode ser feita simplesmente supondo que $\sigma \approx s$. Esta aproximação é um exemplo de *estimativa pontual*, como veremos depois na seção 4.8. Para amostras grandes, s fornece uma boa estimativa de σ, e não haverá problema. Contudo, se a amostra for pequena (o que acontece freqüentemente em áreas como Medicina e Biologia), esta aproximação vai aumentar a incerteza dos resultados; o problema fica mais complicado, e teremos que usar um modelo de distribuição diferente, em lugar do normal (seção 4.7).

Resumo

- Uma *estatística amostral* é uma medida calculada numa amostra (p. ex., a média $\overline{X}$); um *parâmetro* é uma medida que caracteriza uma população (por ex., a média μ).
- A *Inferência* procura tirar conclusões sobre um *parâmetro* a partir da *estatística amostral* correspondente; por exemplo, testar uma hipótese sobre μ a partir de $\overline{X}$.
- A estatística amostral é tratada como uma VAC, cuja distribuição é chamada de *distribuição amostral*.