9

UNIDADE

Estimação de parâmetros

Nesta Unidade você conhecerá e aplicará os conceitos de Estimação de Parâmetros por Ponto e por Intervalo de Média e Proporção, e aprenderá como calcular o tamanho mínimo de amostra necessário para a Estimação por Intervalo.

Estimação por Ponto de Parâmetros

Prezado estudante!

Na Unidade 8 você viu o conceito de Distribuição Amostral e observou a importância do modelo normal. Nesta Unidade você vai aprender como aplicar esses conceitos no primeiro tipo particular de Inferência Estatística, a **Estimação de Parâmetros**: por ponto e por intervalo.

Parâmetros são medidas de síntese de variáveis quantitativas na População que estamos pesquisando. Por ser inviável ou inconveniente pesquisar toda a População coletamos uma amostra para estudá-la. Os resultados da amostra podem ser então usados para fazer afirmações probabilísticas sobre o parâmetro de interesse: definir um intervalo possível para os valores do parâmetro e calcular a probabilidade de que o valor real do parâmetro esteja dentro dele (esta é a Estimação por Intervalo).

Vamos aprender como estimar os parâmetros média de uma variável quantitativa e proporção de um dos valores de uma variável qualitativa. Além disso, você vai ver como é possível definir de forma mais acurada o tamanho mínimo de uma amostra aleatória para estimar média e proporção (para esta última apresentamos uma primeira expressão de cálculo Unidade 2).

Uma vez tendo decidido que modelo probabilístico é mais adequado para representar a variável de interesse na População resta obter os seus parâmetros. Nos estudos feitos com base em amostras é preciso escolher qual das estatísticas da amostra será o melhor estimador para cada parâmetro do modelo.

A Estimação por Ponto consiste em determinar qual será o melhor estimador para o parâmetro de interesse.

Como os parâmetros serão estimados através das estatísticas, estimadores, de uma amostra aleatória, e como para cada amostra aleatória as estatísticas apresentarão diferentes valores, os estimadores também terão valores aleatórios. Em outras palavras um Estimador é uma variável aleatória que pode ter um modelo probabilístico para descrevê-la.

Naturalmente haverá várias estatísticas **T** que poderão ser usadas como estimadores de um parâmetro **θ** qualquer. Como escolher qual das estatísticas será o melhor estimador para o parâmetro?

Há basicamente três critérios para a escolha de um estimador: o estimador precisa ser justo, consistente e eficiente.

Estimação por ponto – tipo de estimação de parâmetros que procura identificar qual é o melhor estimador para um parâmetro populacional a partir das várias estatística amostrais disponíveis, seguindo alguns critérios. Fonte: Barbetta, Reis e Bornia (2008).

1) Um Estimador **T** é um estimador **justo** (não tendencioso) de um parâmetro **θ** quando o valor esperado de **T** é igual ao valor do parâmetro **θ** a ser estimado: **E(T) = θ**

2) Um Estimador T é um estimador **consistente** de um parâmetro **θ** quando além ser um estimador justo a sua variância tende a zero à medida que o tamanho da amostra aleatória aumenta: $\lim_{n\to\infty} V(T) = 0$.

3) Se há dois Estimadores justos de um parâmetro o mais **eficiente** é aquele que apresentar a menor variância.

Conforme foi dito na introdução desta Unidade, estamos interessados em estimar dois parâmetros: média e proporção populacional. Vamos então buscar os estimadores mais apropriados para ambos.

Estimação por ponto dos principais parâmetros

Os principais parâmetros que vamos avaliar aqui são: média de uma variável que segue um modelo normal (ou qualquer modelo se a amostra for suficientemente grande) em uma população (média populacional – μ) e proporção de ocorrência de um dos valores de

uma variável que segue um modelo Binomial em uma população (proporção populacional – π). Em suma escolher quais estatísticas amostrais são mais adequadas para estimar esses parâmetros, usando os critérios definidos acima.

Lembrando dos Exemplos 2, e 3 da Unidade 8, algumas constatações que lá foram feitas passarão a fazer sentido agora.

Vamos supor que houvesse a intenção de estimar a média populacional da variável do Exemplo 2. Qual das estatísticas disponíveis seria o melhor estimador?

Lembrem-se de que após retirar todas as amostras aleatórias possíveis daquela população, calculamos a média de cada amostra, e posteriormente a média dessas médias. Constatou-se que o valor esperado das médias amostrais (média das médias) é igual ao valor da média populacional da variável e a variância das médias amostrais é igual ao valor da variância populacional da variável dividida pelo tamanho da amostra:

$$E(\bar{x}) = \mu \qquad V(\bar{x}) = \frac{\sigma^2}{n}$$

O melhor estimador da média populacional μ é a média amostral $\bar{x}$, pois se trata de um estimador justo e consistente:

- Justo porque o valor esperado da média amostral será a média populacional;
- Consistente porque se o tamanho da amostra **n** tender ao infinito a variância da média amostral (do Estimador) tenderá a zero.

Agora vamos supor que houvesse a intenção de estimar a proporção populacional do valor ■ da variável do Exemplo 3. Qual das estatísticas disponíveis seria o melhor estimador?

Lembrem-se de que após retirar todas as amostras aleatórias possíveis daquela população, calculamos a proporção de ■ em cada amostra, e posteriormente a média dessas proporções. Constatou-se que o valor esperado das proporções amostrais (média das proporções) é igual ao valor da proporção populacional do valor ■ da variável e a variância das proporções amostrais é igual ao valor do produto da proporção populacional do valor ■ da variável pela sua complementar dividida pelo tamanho da amostra:

$$E(p) = \pi \qquad V(p) = \frac{\pi \times (1-\pi)}{n}$$

O melhor estimador da proporção populacional π é a proporção amostral **p**, pois se trata de um estimador **justo** e **consistente**:

- Justo porque o valor esperado da proporção amostral será a proporção populacional; e
- Consistente porque se o tamanho da amostra **n** tender ao infinito a variância da proporção amostral (do Estimador) tenderá a zero.

Poderíamos fazer um procedimento semelhante para estimar outros parâmetros, como, por exemplo, a variância populacional de uma variável. Este procedimento não será demonstrado, mas o melhor estimador da variância populacional será a variância amostral se for usado **n – 1** no denominador da expressão de cálculo. Somente assim a variância amostral será um estimador justo (não viciado) da variância populacional.

Como o desvio padrão é a raiz quadrada da variância é comum estimar o desvio padrão populacional extraindo a raiz quadrada da variância amostral.

O problema da Estimação por Ponto é que geralmente só dispomos de uma amostra aleatória. Intuitivamente, qual será a probabilidade de que a média ou proporção amostral, de uma amostra aleatória, coincida exatamente com o valor do parâmetro? É como pescar usando uma lança de bambu... É preciso muita habilidade para pegar o peixe... Mas, se você puder usar uma rede, ficará bem mais fácil. Essa "rede" é a Estimação por Intervalo.

Estimação por Intervalo de Parâmetros

Geralmente uma inferência estatística é feita com base em uma única amostra: na maior parte dos casos é totalmente inviável retirar todas as amostras possíveis de uma determinada população.

Intuitivamente percebemos que as estatísticas calculadas nessa única amostra, mesmo sendo os melhores estimadores para os parâmetros de interesse, terão uma probabilidade infinitesimal de coincidir exatamente com os valores reais dos parâmetros. Então a Estimação por Ponto dos parâmetros é insuficiente, e as estimativas assim obtidas servirão apenas como referência para a Estimação por Intervalo.

A Estimação por Intervalo consiste em colocar um Intervalo de Confiança (I.C.) em torno da estimativa obtida através da Estimação por Ponto.

O Intervalo de Confiança terá uma certa probabilidade chamada de Nível de confiança (que costuma ser simbolizado como $\mathbf{1 - \alpha}$) de conter o valor real do parâmetro e a probabilidade de que esta faixa realmente contenha o valor real do parâmetro. A probabilidade de que o Intervalo de Confiança não contenha o valor real do parâmetro é chamada de Nível de Significância ($\boldsymbol{\alpha}$), e o valor desta probabilidade será o complementar do Nível de Confiança. É comum definir o Nível de Significância como uma probabilidade máxima de erro, um risco máximo admissível.

A determinação do Intervalo de Confiança para um determinado parâmetro resume-se basicamente a definir o Limite Inferior e o Limite Superior do intervalo, supondo um determinado Nível de Confiança ou Significância.

A definição dos limites dependerá também da distribuição amostral da estatística usada como referência para o intervalo e do tamanho da amostra utilizada.

Para os dois parâmetros em que temos maior interesse (média populacional $\boldsymbol{\mu}$ e proporção populacional $\boldsymbol{\pi}$) a distribuição amostral dos estimadores (média amostral $\bar{x}$ e proporção amostral $\mathbf{p}$, respectivamente) pode ser aproximada por uma distribuição normal: o Intervalo de Confiança será então simétrico em relação ao valor calculado da estimativa (média ou proporção amostral), com base na amostra aleatória coletada (Figura 68):

fazer uma Estimação por Intervalo de um parâmetro é efetuar uma afirmação probabilística sobre este parâmetro, indicando uma faixa de possíveis valores.

Intervalo de confiança – faixa de valores da estatística usada como estimador, dentro da qual há uma probabilidade conhecida de que o verdadeiro valor do parâmetro esteja. Sinônimo de estimação por intervalo. Fonte: Barbetta, Reis e Bornia (2008).

Nível de Significância – complementar do nível de confiança, a probabilidade de que o intervalo de confiança não contenha o valor real do parâmetro. Probabilidade de erro espera-se que seja um valor baixo, de no máximo 10%. Fonte: Barbetta, Reis e Bornia (2008).

Nível de confiança – probabilidade de que o intervalo de confiança contenha o valor real do parâmetro a estimar, espera-se que seja um valor alto, de no mínimo 90%. Fonte: Moore, McCabe, Duckworth e Sclove (2006).

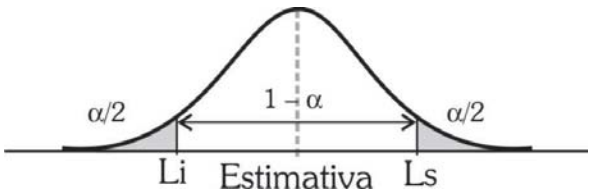

Figura 68: Intervalo de Confiança para um modelo normal.
Fonte: elaborada pelo autor.

Onde: $\mathbf{L_i}$ é o limite inferior e $\mathbf{L_s}$ é o limite superior do Intervalo de Confiança; $\mathbf{1 - \alpha}$ é o Nível de Confiança estabelecido, observando que o valor do Nível de Significância $\boldsymbol{\alpha}$ é dividido igualmente entre os valores abaixo de $\mathbf{L_i}$ e acima de $\mathbf{L_s}$.

Para obter os limites em função do Nível de Confiança devemos utilizar a distribuição normal padrão (variável **Z** com média zero e variância um): fixar um certo valor de probabilidade, obter o valor de Z correspondente, e substituir o valor em **Z** = (**x** – "média") "desvio padrão", para obter o valor x (valor correspondente ao valor de **Z** para a probabilidade fixada). Observe a Figura 69:

Foram colocados entre aspas porque os valores dependerão dos parâmetros sob análise e de outros fatores.

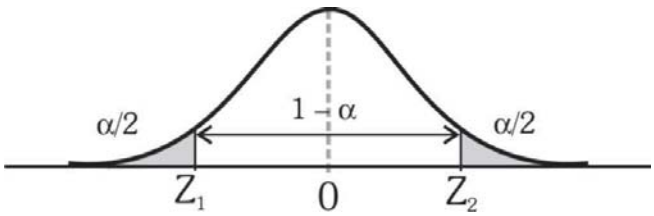

Figura 69: Intervalo de Confiança para a distribuição normal padrão.
Fonte: elaborada pelo autor.

O limite $\mathbf{L_i}$ (inferior) corresponde a $\mathbf{Z_1}$ e o limite $\mathbf{L_s}$ (superior) corresponde a $\mathbf{Z_2}$. O ponto central 0 (zero) corresponde ao valor calculado da Estimativa. Como a variável **Z** tem distribuição normal com média igual a zero (lembrando que a distribuição normal é simétrica em relação à média) os valores de $\mathbf{Z_1}$ e $\mathbf{Z_2}$ serão iguais em módulo ($\mathbf{Z_1}$ será negativo e $\mathbf{Z_2}$ positivo):

$\mathbf{Z_1}$ será um valor de **Z** tal que $P(Z \leq Z_1) = \frac{\alpha}{2}$, e $\mathbf{Z_2}$ será um valor tal que $P(Z \leq Z_2) = 1 - \frac{\alpha}{2}$.

Então, obteremos os valores dos limites através das expressões:

$Z_1 = (L_i - \text{“média”})/ \text{“desvio padrão”} \Rightarrow L_i = \text{“média”} + Z_1 \times \text{“desvio padrão”}.$

$Z_2 = (L_s - \text{“média”})/ \text{“desvio padrão”} \Rightarrow L_s = \text{“média”} + Z_2 \times \text{“desvio padrão”}.$

Como $\mathbf{Z_1 = - Z_2}$, podemos substituir:

$L_i = \text{“média”} - Z_2 \times \text{“desvio padrão”}.$

$L_s = \text{“média”} + Z_2 \times \text{“desvio padrão”}.$

E este valor $\mathbf{Z_2}$ costuma ser chamado de $\mathbf{Z_{crítico}}$, porque corresponde aos limites do intervalo:

$L_i = \text{“média”} - Z_{crítico} \times \text{“desvio padrão”}.$

$L_s = \text{“média”} + Z_{crítico} \times \text{“desvio padrão”}.$

Reparem que o mesmo valor é somado, e subtraído da “média”. Esse valor é chamado de semi-intervalo ou precisão do intervalo, ou margem de erro, $\mathbf{e_0}$:

$e_0 = Z_{crítico} \times \text{“desvio padrão”}.$

Resta agora definir corretamente o valor da “média” e do “desvio padrão” para cada um dos parâmetros em que estamos interessados (média e proporção populacional). Com base nas conclusões obtidas na Estimação por Ponto isso será simples. Contudo, há alguns outros aspectos que precisarão ser esmiuçados.

Estimação por Intervalo da Média Populacional

Lembrando das expressões anteriores:

$L_i = \text{“média”} - Z_{crítico} \times \text{“desvio padrão”} = \text{“média”} - e_0.$

$L_s = \text{“média”} + Z_{crítico} \times \text{“desvio padrão”} = \text{“média”} + e_0.$

Neste caso a “média” será a média amostral $\bar{x}$ (ou mais precisamente o seu valor):

$$\mathbf{P(\bar{x} - e_0 \leq \mu \leq \bar{x} + e_0) = 1 - \alpha}$$

O valor de $\mathbf{e_0}$ dependerá de outros aspectos.

a) Se a variância populacional σ^2 da variável (cuja média populacional queremos estimar) for conhecida.

Neste caso a variância amostral da média poderá ser calculada através da expressão:

$V(\bar{x}) = \frac{\sigma^2}{n}$, e, por conseguinte, o "desvio padrão" será $= \frac{\sigma}{\sqrt{n}}$

E e_0 será: $e_0 = Z_{crítico} \times \frac{\sigma}{\sqrt{n}}$

Bastará então fixar o Nível de Confiança (ou de Significância) para obter $Z_{crítico}$ através da Tabela disponível no Ambiente Virtual e calcular e_0.

b) Se a variância populacional σ^2 da variável for desconhecida.

Naturalmente este é o caso mais encontrado na prática. Como se deve proceder? Dependerá do tamanho da amostra.

b.1 – Grandes amostras (mais de 30 elementos).

Nestes casos procede-se como no item anterior, apenas fazendo com que $\sigma = s$, ou seja, considerando que o desvio padrão da variável na população é igual ao desvio padrão da variável na amostra (suposição razoável para grandes amostras).

b.2 – Pequenas amostras (até 30 elementos).

Nestes casos a aproximação do item b.1 não será viável. Terá que ser feita uma correção na distribuição normal padrão (**Z**) através da distribuição **t de Student** que estudamos na Unidade 6.

Quando a variância populacional da variável é desconhecida e a amostra tem até 30 elementos substitui-se σ por **s** e **Z** por t_{n-1} em todas as expressões para determinação dos limites do intervalo de confiança, obtendo:

L_i = "média" – $t_{n-1,\ crítico} \times$ "desvio padrão" = "média" – e_0

L_s = "média" + $t_{n-1,\ crítico} \times$ "desvio padrão" = "média" + e_0

E e_0 será: $e_0 = t_{n-1,crítico} \times \frac{s}{\sqrt{n}}$

Os valores de $t_{n-1,crítico}$ podem ser obtidos de forma semelhante aos de $Z_{crítico}$, definindo o Nível de Confiança (ou de Significância), mas precisam também da definição do número de graus de liberdade (**n – 1**): tendo estes valores basta procurar o valor da Tabela 2 do Ambiente Virtual ou em um programa computacional.

Se o tamanho da amostra (n) for superior a 5% do tamanho da população (N) os valores de e0 precisam ser corrigidos. Caso contrário os limites dos intervalos não serão acusados. A correção é mostrada na equação a seguir:

$$e_{0_{corrigido}} = e_0 \times \sqrt{\frac{N-n}{N-1}}$$

Vamos ver um exemplo.

Neste primeiro exemplo retirou-se uma amostra aleatória de quatro elementos de uma produção de cortes bovinos no intuito de estimar a média do peso do corte. Obteve-se média de 8,2 kg e desvio padrão de 0,4 kg. Supondo população normal.

Determinar um intervalo de confiança para a média populacional com 1% de significância.

O parâmetro de interesse é a média populacional μ do peso do corte.

Adotou-se um nível de significância de 1%, então $\alpha = 0{,}01$ e $1 - \alpha = 0{,}99$.

Este valor pode ser arbitrado pelo usuário ou pode ser uma exigência do problema sob análise, ou até mesmo uma exigência legal. Os níveis de significância mais comuns são de 1%, 5% ou mesmo 10%.

As estatísticas disponíveis são: **média amostral** = 8,2 kg **s** = 0,4 kg **n** = 4 elementos.

Definição da variável de teste: como a variância populacional é DESCONHECIDA, e o tamanho da amostra é menor do que 30 elementos, não obstante a população ter distribuição normal, a distribuição amostral da média será **t** de *Student*, e a variável de teste será $\mathbf{t_{n-1}}$.

Encontrar o valor de $\mathbf{t_{n-1,crítico}}$: como o Intervalo de Confiança para a média é bilateral, teremos uma situação semelhante à da Figura 70:

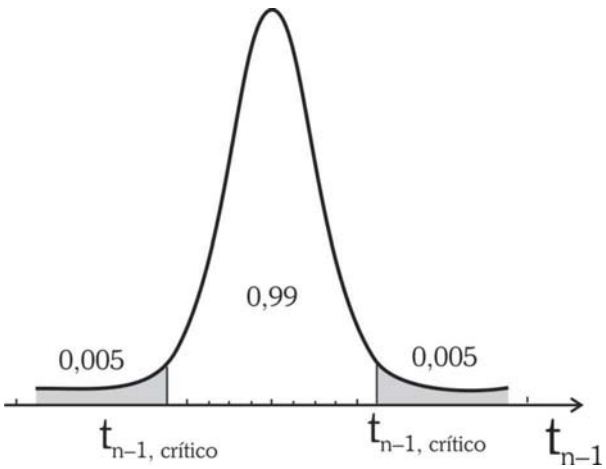

Figura 70: Distribuição t de *Student* para 99% de confiança.
Fonte: elaborada pelo autor.

Para encontrar o valor crítico devemos procurar na tabela da distribuição de *Student*, na linha correspondente a **n–1** graus de liberdade, ou seja, em 4 – 1 = 3 graus de liberdade. O valor da probabilidade pode ser visto na Figura acima: os valores críticos serão $\mathbf{t_{3;0,005}}$ e $\mathbf{t_{3;0,995}}$ os quais serão iguais em módulo. E o valor de $\mathbf{t_{n-1,crítico}}$ será igual a **5,84** (em módulo).

Determinam-se os limites do intervalo, através da expressão abaixo (cujo resultado será somado e subtraído da média amostral) para determinar os limites do intervalo:

$$e_0 = \frac{t_{n-1,crítico} * s}{\sqrt{n}} = \frac{5{,}84 * 0{,}4}{\sqrt{4}} = 1{,}168 \text{ kg}.$$

$L_I = \bar{x} - e_0 = 8{,}2 - 1{,}168 = 7{,}032$ kg.
$L_S = \bar{x} + e_0 = 8{,}2 + 1{,}168 = 9{,}368$ kg.

Então o intervalo de 99% de confiança para a média populacional da dimensão é [7,032;9,368] kg. Interpretação: há 99% de probabilidade de que a verdadeira média populacional do peso de corte esteja entre 7,032 e 9,368 kg.

Estimação por Intervalo da Proporção Populacional

Anteriormente declaramos que o melhor estimador para a proporção populacional $\boldsymbol{\pi}$ é a proporção amostral **p**. E que esta proporção amostral teria média igual a $\boldsymbol{\pi}$ e variância igual a $[\boldsymbol{\pi} \times (1 - \boldsymbol{\pi})]/\mathbf{n}$ onde **n** é o tamanho da amostra aleatória. A distribuição da proporção amostral p é binomial, e sabemos que a distribuição binomial pode ser aproximada por uma normal se algumas condições forem satisfeitas:

$$\text{Se } n \times \boldsymbol{\pi} \geq 5 \text{ E } \mathbf{n} \times (1 - \boldsymbol{\pi}) \geq 5.$$

Ora, se $\boldsymbol{\pi}$ fosse conhecido não estaríamos aqui nos preocupando com a sua Estimação por Intervalo, assim vamos verificar se é possível aproximar a distribuição binomial de **p** por uma normal se: $\mathbf{n} \times \mathbf{p} \geq 5$ E $\mathbf{n} \times (1 - \mathbf{p}) \geq 5$, ou seja usando o próprio valor da proporção amostral observada (trata-se de uma aproximação razoável).

Se e somente se estas duas condições forem satisfeitas poderemos usar as expressões abaixo (lembrando das expressões anteriores):

L_i = "média" – $Z_{crítico}$ × "desvio padrão" = "média" – e_0.

L_s = "média" + $Z_{crítico}$ × "desvio padrão" = "média" + e_0.

Neste caso a "média" será a proporção amostral (ou mais precisamente o seu valor):

$$\mathbf{P(p - e_0 \leq \mu \leq p + e_0) = 1 - \alpha}$$

E o valor do "desvio padrão" será igual a $\sqrt{\frac{\pi \times (1-\pi)}{n}}$. Novamente, como $\boldsymbol{\pi}$ é desconhecido, usaremos a proporção amostral **p** como aproximação.

Então $\mathbf{e_0}$ será: $e_0 = Z_{crítico} \times \sqrt{\frac{p \times (1-p)}{n}}$.

Bastará então fixar o Nível de Confiança (ou de Significância), $\mathbf{Z_{crítico}}$ e calcular $\mathbf{e_0}$.

Novamente, precisamos corrigir o valor de e0 para o caso de população finita:

$$e_{0_{corrigido}} = e_0 \times \sqrt{\frac{N-n}{N-1}}.$$

Em suma a Estimação por Intervalo da média e da proporção populacional consiste basicamente em calcular a amplitude do semi-intervalo (o $\mathbf{e_0}$), de acordo com as condições do problema sob análise.

- Para a média, observar se é viável considerar que a distribuição da variável na população é normal, ou que a amostra seja suficientemente grande para que a distribuição das médias amostrais possa ser considerada normal;
- Se isso for verificado, identificar se a variância populacional da variável é conhecida: caso seja deverá ser usada a variável **Z** da distribuição normal padrão, para qualquer tamanho de amostra;
- Se variância populacional da variável é desconhecida há duas possibilidades: para amostras com mais de 30 elementos usar a variável **Z**, e fazer a variância populacional igual à variância amostral da variável; se a amostra tem até 30 elementos usar a variável $\mathbf{t_{n-1}}$ da distribuição de *Student*; e

- Para a proporção, observar se é possível fazer a aproximação pela distribuição normal.

Vamos ver um exemplo.

No exemplo 2, retirou-se uma amostra aleatória de 1000 peças de um lote. Verificou-se que 35 eram defeituosas.

Determinar um intervalo de confiança de 95% para a proporção peças defeituosas no lote.

O parâmetro de interesse é a proporção populacional $\boldsymbol{\pi}$ de peças defeituosas.

Adotou-se um nível de significância de 5%, então $\boldsymbol{\alpha} = 0{,}05$ e $\mathbf{1 - \alpha} = 0{,}95$

As estatísticas são: proporção amostral de peças defeituosas $\mathbf{p} = 35/1000$ $\mathbf{n} = 1000$ elementos.

Definição da variável de teste: precisamos verificar se é possível fazer a aproximação pela normal, então $\mathbf{n} \times \mathbf{p} = 1000 \times 0{,}035 = 35 > 5$ e $\mathbf{n} \times (\mathbf{1 - p}) = 1000 \times 0{,}965 = 965 > 5$. Como ambos os produtos satisfazem as condições para a aproximação podemos usar a variável **Z** da distribuição normal padrão.

Encontrar o valor de $\mathbf{Z_{crítico}}$: como o Intervalo de Confiança para a média é bilateral, teremos uma situação semelhante à da figura abaixo:

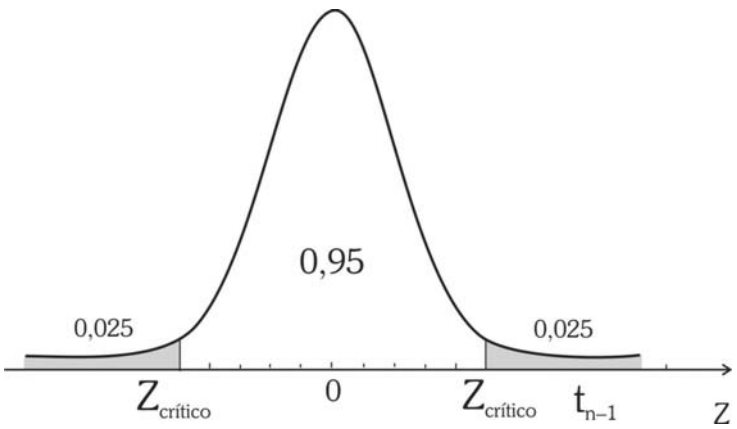

Figura 71: Distribuição normal padrão para 95% de confiança.
Fonte: elaborada pelo autor.

Para encontrar o valor crítico devemos procurar na tabela da distribuição normal padrão pela probabilidade 0,975 (0,95+0,025). O valor da probabilidade pode ser visto na Figura 71 acima: os valores críticos serão $\mathbf{Z_{0,025}}$ e $\mathbf{Z_{0,975}}$ os quais serão iguais em módulo. E o valor de $\mathbf{Z_{crítico}}$ será igual a 1,96 (em módulo).

Passamos agora à determinação dos limites do intervalo, através da expressão abaixo, cujo resultado será somado e subtraído da proporção amostral de peças defeituosas, para determinar os limites do intervalo:

$$e_0 = Z_{crítico} \times \sqrt{\frac{p \times (1-p)}{n}} = 1{,}96 \times \sqrt{\frac{0{,}035 \times 0{,}965}{1000}} = 0{,}0114$$

$L_I = p - e_0 = 0{,}035 - 0{,}0114 = 0{,}0236$

$L_S = p + e_0 = 0{,}035 + 0{,}0114 = 0{,}0464$

Então, o intervalo de 95% de confiança para a proporção populacional de peças defeituosas é [2,36%;4,64%]. Interpretação: há 95% de probabilidade de que a verdadeira proporção populacional de plantas atacadas pelo fungo esteja entre 2,36% e 4,64%.

Tamanho mínimo de amostra para Estimação por Intervalo

Como observado nos itens anteriores, a determinação dos limites de um Intervalo de Confiança (determinação do $\mathbf{e_0}$) depende do tamanho da amostra aleatória coletada, além do Nível de Confiança e da distribuição amostral do estimador utilizado. Nada podemos fazer quanto à distribuição amostral do estimador, o Nível de Confiança nós podemos controlar, seria interessante definir então uma **precisão** (um valor para $\mathbf{e_0}$) para o Intervalo de Confiança: é muito comum querermos estabelecer previamente qual será a faixa de variação de um determinado parâmetro, com uma certa confiabilidade.

Contudo, para um mesmo tamanho de amostra:

- Se aumentarmos o Nível de Confiança (reduzirmos o Nível de Significância) teremos um valor crítico maior, o que aumentará o valor de $\mathbf{e_0}$, resultando em um Intervalo de Confiança mais "largo", com menor precisão.
- Se resolvermos aumentar a precisão (menor valor de $\mathbf{e_0}$), obter um Intervalo de Confiança mais "estreito", teremos uma queda no Nível de Confiança.

A solução para o dilema acima é obter um **tamanho mínimo de amostra** capaz de atender simultaneamente ao Nível de Confiança (ou de Significância) e à precisão ($\mathbf{e_0}$) especificados. Como as expressões de $\mathbf{e_0}$ são em função do tamanho de amostra (**n**), seria razoável pensar em reordená-las de forma a fazer com que o tamanho de amostra seja função do Nível de Confiança e da precisão ($\mathbf{e_0}$).

Tamanho mínimo de amostra para Estimação por Intervalo da Média Populacional

a) Variância populacional conhecida:

$$e_0 = Z_{crítico} \times \frac{\sigma}{\sqrt{n}} \qquad \text{isolando } \mathbf{n}\text{: } n = \left(\frac{Z_{crítico} \times \sigma}{e_0} \right)^2.$$

Neste caso basta especificar o valor de $\mathbf{e_0}$ (na **mesma unidade** do desvio padrão populacional σ), e o Nível de Confiança (que será usado para encontrar o $\mathbf{Z_{crítico}}$) e calcular o tamanho mínimo de amostra.

b) Variância populacional desconhecida

$$e_0 = t_{n-1,crítico} \times \frac{s}{\sqrt{n}} \qquad \text{isolando } \mathbf{n}\text{: } n = \left(\frac{t_{n-1,crítico} \times s}{e_0} \right)^2.$$

O procedimento neste caso seria semelhante exceto por um pequeno problema: "se estamos calculando o tamanho da amostra como podemos conhecer **n – 1** e o desvio padrão amostral **s**?"

Quando a variância populacional da variável é desconhecida o usual é retirar uma amostra piloto com um tamanho **n*** arbitrário. A partir dos resultados desta amostra são calculadas as estatísticas (entre elas o desvio padrão amostral **s**) que são substituídas na expressão acima.

Amostra piloto – amostra teste, de tamanho arbitrado pelo pesquisador, a partir da qual são calculadas estatísticas necessárias para a determinação do tamanho mínimo de amostra. Fonte: Costa Neto (2002).

Se $\mathbf{n} \leq \mathbf{n^*}$ então a amostra piloto é suficiente para o Nível de Confiança e a precisão exigidos.

Se $\mathbf{n} > \mathbf{n^*}$ então a amostra piloto é insuficiente para o Nível de Confiança e a precisão exigidas, sendo então necessário retornar à população e retirar os elementos necessários para completar o tamanho mínimo de amostra. O processo continua até que a amostra seja considerada suficiente.

Conforme visto na Unidade 2, se o tamanho da população for conhecido é recomendável corrigir o tamanho da amostra obtida, seja

para o intervalo de confiança de média ou proporção, através da seguinte fórmula:

$$n_{corrigido} = \frac{N \times n}{N + n}$$ onde N é o tamanho da população

Assim procedendo, evitamos o inconveniente de obter um tamanho de amostra superior ao tamanho da população, o que pode ocorrer se N não for muito grande.

Considerem, neste exemplo 3, os dados do Exemplo 1. Para estimar a média, com 1% de significância e precisão de 0,2 kg, esta amostra é suficiente

Como a variância populacional é desconhecida, e o tamanho da amostra é menor do que 30 elementos, não obstante a população ter distribuição normal, a distribuição amostral da média será **t** de *Student*, e a variável de teste será $\mathbf{t_{n-1}}$. Assim será usada a seguinte expressão para calcular o tamanho mínimo de amostra para a estimação por intervalo da média populacional.

$$n = \left(\frac{t_{n-1,crítico} \times s}{e_0} \right)^2$$

O nível de significância é o mesmo do item a. Sendo assim, o valor crítico continuará sendo o mesmo: $\mathbf{t_{n-1,crítico}} = 5{,}84$. O desvio padrão amostral vale 0,4 kg, e o valor de $\mathbf{e_0}$, a precisão foi fixado em 0,2 kg. Basta então substituir os valores na expressão:

$$n = \left(\frac{t_{n-1,crítico} \times s}{e_0} \right)^2 = \left(\frac{5{,}84 \times 0{,}4}{0{,}2} \right)^2 = 136{,}42 \cong 137 \text{ elementos.}$$

Conclui-se que a amostra retirada é insuficiente, pois é menor do que o valor calculado acima.

Tamanho mínimo de amostra para Estimação por Intervalo da Proporção Populacional

Para a proporção populacional teremos:

$$e_0 = Z_{critico} \times \sqrt{\frac{p \times (1-p)}{n}}$$ isolando n: $$n = \left(\frac{Z_{critico}}{e_0} \right)^2 \times p \times (1-p)$$

É necessário especificar o Nível de Confiança (ou de Significância) que será usado para encontrar o $\mathbf{Z_{crítico}}$, e o valor de $\mathbf{e_0}$ (tomando o cuidado de que tanto $\mathbf{e_0}$ quanto **p** e **1– p** estejam **todos** como proporções adimensionais ou como percentuais) para que seja possível calcular o valor do tamanho mínimo de amostra.

Da mesma forma que no caso da Estimação da média quando a variância populacional é desconhecida teremos que recorrer à uma amostra piloto. No cálculo do tamanho mínimo de amostra para a Estimação por Intervalo da proporção populacional há, porém, uma solução alternativa: utiliza-se uma estimativa exagerada da amostra, supondo o máximo valor possível para o produto $\mathbf{p} \times (\mathbf{1} - \mathbf{p})$, que ocorrerá quando ambas as proporções forem iguais a 0,5 (50%).

Esta solução somente é usada quando a natureza da pesquisa é tal que não é possível retirar uma amostra piloto: a retirada de uma amostra piloto e a eventual retirada de novos elementos da população poderiam prejudicar muito o resultado da pesquisa. Paga-se, então, o preço de ter uma amostra substancialmente maior do que talvez fosse necessário.

Conforme visto na Unidade 2, se o tamanho da população for conhecido é recomendável corrigir o tamanho da amostra obtida, seja para o intervalo de confiança de média ou proporção, através da seguinte fórmula:

$$n_{corrigido} = \frac{N \times n}{N + n}$$, onde N é o tamanho da população.

Assim procedendo, evitamos o inconveniente de obter um tamanho de amostra superior ao tamanho da população, o que pode ocorrer se N não for muito grande.

Neste quarto exemplo considere o caso do Exemplo 2. Supondo 99% de confiança e precisão de 1%, esta amostra é suficiente para estimar a proporção populacional

De acordo com o Exemplo 2 é possível utilizar a aproximação pela distribuição normal. A expressão para o cálculo do tamanho mínimo de amostra para a proporção populacional será:

$$n = \left(\frac{Z_{crítico}}{e_0} \right)^2 \times p \times (1 - p).$$

Os valores de **p** e **1 – p** já são conhecidos:

p = 0,035 **1 – p** = 0,965

O nível de confiança exigido é de 99%: para encontrar o valor crítico devemos procurar na tabela da distribuição normal padrão pela probabilidade 0,995 (0,99+0,005); os valores críticos serão $\mathbf{Z_{0,005}}$ e $\mathbf{Z_{0,995}}$ os quais serão iguais em módulo. E o valor de $\mathbf{Z_{crítico}}$ será igual a 2,58 (em módulo).

A precisão foi fixada em 1% (0,01). Substituindo os valores na expressão acima:

$$n = \left(\frac{Z_{crítico}}{e_0}\right)^2 \times p \times (1-p) = \left(\frac{2{,}58}{0{,}01}\right)^2 \times 0{,}035 \times 0{,}965 = 2.248{,}14 \cong 2.249$$

Observe que o tamanho mínimo de amostra necessário para atender a 99% de confiança e precisão de 1% deveria ser de 2.249 elementos. Como a amostra coletada possui apenas 1.000 elementos ela é insuficiente para a confiança e precisão exigidas. Recomenda-se o retorno à população para a retirada aleatória de mais 1.249 peças.

Depois de tudo que estudamos você já pode acompanhar atentamente os resultados das pesquisas de opinião veiculadas na mídia. Apenas mais um pequeno adendo.

"Empate técnico"

Estamos acostumados a ouvir declarações do tipo "os candidatos A e B estão tecnicamente empatados na preferência eleitoral". O que significa isso? Geralmente as pesquisas de opinião eleitoral consistem em obter as proporções de entrevistados que declara votar neste ou naquele candidato, naquele momento. Posteriormente as proporções são generalizadas estatisticamente para a população, através do cálculo de intervalos de confiança para as proporções de cada candidato. Se os intervalos de confiança das proporções de dois ou mais candidatos apresentam grandes superposições declara-se que há um "empate técnico": as diferenças entre eles devem-se provavelmente ao acaso, e para todos os fins estão em condições virtualmente iguais, naquele momento.

Neste exemplo 5, imagine que uma pesquisa de opinião eleitoral apresentasse os seguintes resultados (intervalos de confiança para a proporção que declara votar no candidato) sobre a prefeitura do município de Tapioca. Quais candidatos estão tecnicamente empatados (Quadro 22)?

Opinião	Limite inferior %	Limite superior %
Godofredo Astrogildo	31%	37%
Filismino Arquibaldo	14%	20%
Urraca Hermengarda	13%	19%
Salustiano Quintanilha	22%	28%
Indecisos	11%	17%

Quadro 22: Resultados de uma pesquisa eleitoral municipal (fictícia).
Fonte: elaborado pelo autor.

Filismino e Urraca estão tecnicamente empatados, pois seus intervalos de confiança apresentam grande sobreposição. Godofredo está muito na frente, pois o limite inferior de seu intervalo é maior do que o limite superior de Salustiano, que está em segundo lugar. É importante ressaltar que o número de indecisos é razoável, variando de 11 a 17%, quando eles se decidirem poderão mudar completamente o quadro da eleição, ou garantir a vitória folgada de Godofredo.

Saiba mais

Sobre propriedades e características desejáveis de um estimador, BARBETTA, Pedro A.; REIS, Marcelo M.; BORNIA, Antonio C. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2008, Capítulo 7.

Sobre estimadores e intervalos de confiança para variância, TRIOLA, Mario. *Introdução à Estatística*. Rio de Janeiro: LTC, 1999, Capítulo 6.

Para entender melhor o conceito de distribuição amostral e sua relação com estimação de parâmetros, veja o arquivo Estima.xls, e suas instruções, no ambiente virtual;

Sobre a utilização do Microsoft Excel para realizar estimação por intervalo, LEVINE, David M.; STEPHAN, David; KREHBIEL, Timothy C.; BERENSON, Mark L. *Estatística*: Teoria e Aplicações – Usando Microsoft Excel em Português. 5. ed. Rio de Janeiro: LTC, 200, Capítulo 6.

Resumindo

O resumo desta Unidade está mostrado na Figura 72:

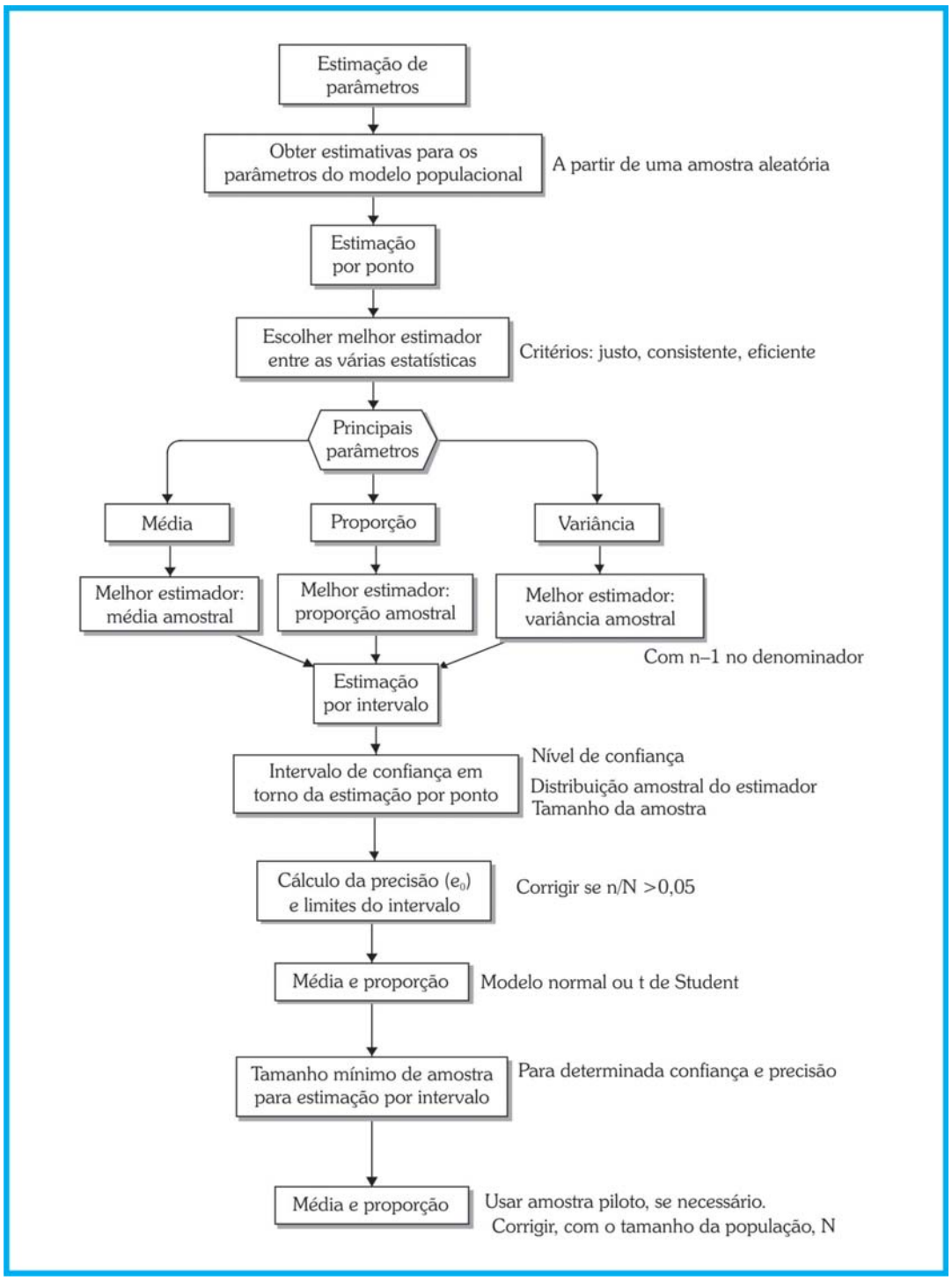

Figura 72: Resumo da Unidade 9.
Fonte: elaborado pelo autor.

Vimos nesta Unidade sobre os conceitos de Estimação de Parâmetros. Aprendemos a estimar os parâmetros média de uma variável quantitativa e proporção de um dos valores de uma variável qualitativa. Além de definir o tamanho mínimo de uma amostra aleatória para estimar média e proporção. Veremos mais sobre esse assunto na última Unidade deste livro. Estamos próximos do final do nosso material e é de suma importância a continuidade da interação com seus colegas e professor. Não deixe de ver as tabelas indicadas no livro e disponíveis no Ambiente Virtual de Ensino-Aprendizagem e de realizar a atividade de aprendizagem.

Atividades de aprendizagem

1) Buscando melhorar a qualidade do serviço, uma empresa estuda o tempo de atraso na entrega dos pedidos recebidos. Supondo que o tempo de atraso se encontra normalmente distribuído, e conhecendo o tempo de atraso dos últimos 20 pedidos, descritos abaixo (em dias), determine:

5 1 0 3 6 10 2 3 4 1 5 3 1 6 6 9 0 0 1 0

a) Estime o atraso médio na entrega dos pedidos com confiança de 90%.

b) Se fosse conhecido que a população possui desvio padrão igual a 2 dias, como ficaria a resposta do item a)?

c) Para a situação do item a (variância populacional desconhecida), o tamanho da amostra é suficiente, se é necessária uma precisão de 0,5 dias, para o mesmo nível de confiança?

2) A satisfação da população em relação a determinado governo foi pesquisada através de uma amostra com a opinião de 1.000 habitantes do estado. Destes, 585 se declararam insatisfeitas com a administração estadual. Admitindo-se um nível de significância de 5%, solucione os itens abaixo.

   a) Estime o percentual da população que está insatisfeita com a administração estadual.

   b) Qual o tamanho da amostra necessária para a estimação se a empresa responsável pela pesquisa estipulou uma folga máxima de 2,5%?

3) Os índices apresentados pelos alunos do curso de Economia e de Administração estão sendo questionados pelos alunos, no sentido de definirem se há diferença entre os cursos. Para tanto foram analisados os índices de 10 alunos de cada curso, escolhidos aleatoriamente dentre os regularmente matriculados e anotados seus valores, onde se obteve:

Economia	– média 7,3	desvio padrão 2,6
Administração	– média 7,1	desvio padrão 3,1

   a) Estime os valores médios dos índices de cada curso com 95% de confiança.

   b) Para o mesmo nível de confiança de a. Será que 10 alunos é uma amostra suficiente, em ambos os cursos, para estimar seus índices médios, com uma precisão igual a 1?

4) O CRA de SC está conduzindo uma pesquisa sobre a opinião dos acadêmicos de administração sobre seus respectivos cursos. Suspeita-se que haja diferença entre as proporções de satisfeitos de instituições públicas e privadas: os acadêmicos das públicas seriam mais satisfeitos. Para avaliar esta suposição foi conduzida uma pesquisa por amostragem, entrevistando alunos de duas instituições públicas, SHUFSC e GASE, e de três privadas, PATÁPIO de SÁ, UNIMALI e UNILUS. Os resultados estão na tabela a seguir:

MEDIDAS	UNIVERSIDADES				
	SHUFSC	GASE	PATÁPIO	UNIMALI	UNILUS
n	120	165	185	194	189
p	0,55	0,48	0,32	0,49	0,25
N (população)	890	900	1500	1200	1800

Usando 1% de significância responda os itens a seguir:

a) Estime a proporção populacional de satisfeitos com o seu curso, em cada universidade*.

b) Para uma margem de erro de 2% qual deveria ser o tamanho de amostra para estimar a proporção de satisfeitos em cada universidade?*