1 Inferência Estatística - Teoria da Estimação

1.1 Introdução

Neste capítulo abordaremos situações em que o interesse está em obter informações da população a partir dos resultados de uma amostra. Como exemplo, consideremos uma indústria de produtos de cabelo que pretende investigar a aceitação entre as mulheres de seu novo produto tonalizante. Para tanto, selecionamos uma amostra aleatória de mulheres que utilizaram o produto e verificamos qual é o percentual de aprovação desse produto na amostra. Outro exemplo trata-se de um psiquiatra interessado em determinar o tempo de reação de um medicamento anti-depressivo em crianças. Uma amostra aleatória de crianças que utilizaram o medicamento é obtida e analisamos o tempo médio de reação. Nestes dois exemplos, deseja-se determinar o valor de um parâmetro por meio da avaliação de uma amostra.

A seguir vamos definir alguns conceitos básicos de inferência estatística.

- Parâmetro: é uma medida numérica, em geral desconhecida, que descreve uma característica de interesse da população. São representados, geralmente, por letras gregas tais como, μ (média populacional) e σ (desvio-padrão populacional), entre outros. Neste texto, usaremos a letra P para representar a proporção populacional.

- Estatística: é qualquer valor calculado a partir dos dados amostrais. Por exemplo, $\bar{X}$ (média amostral), S (desvio-padrão amostral), $\hat{P}$ (proporção amostral), etc. A estatística é uma variável aleatória, no sentido que (a) é uma quantidade incerta (antes de obter a amostra não sabemos seu valor) e (b) seu valor varia de amostra para amostra. É claro que, quando uma amostra é selecionada, e uma estatística é calculada, torna-se então uma constante, ou seja, é o resultado de uma variável aleatória.

- Estimador e Estimativa: uma estatística destinada a estimar um parâmetro populacional é chamada estimador. Dada uma amostra, o valor assumido pelo estimador é chamado de estimativa ou valor estimado do parâmetro. As estimativas obtidas por meio da estatística variam de acordo com a amostra selecionada.

Para diferenciar estimador de estimativa, vamos representar os estimadores por letras maiúsculas e as correspondentes estimativas por letras minúsculas. Por exemplo, um estimador da média populacional, μ, é a média amostral $\bar{X}$. Dada uma amostra da população, a estimativa de μ é $\bar{x}$.

Os procedimentos de inferência estatística compreendem duas metodologias. Uma é chamada de estimação, na qual nós usamos o resultado amostral para estimar o valor desconhecido do parâmetro; a outra, é conhecida como teste de hipóteses, em que nós usamos o resultado amostral para avaliar se uma afirmação sobre

o parâmetro (uma hipótese) é sustentável ou não. Teoria de estimação é o assunto principal deste capítulo e teste de hipóteses será retomado no próximo capítulo.

É importante estudar as propriedades do estimador, para avalia-lo, ou seja, para poder responder a pergunta: Será que é um bom estimador para o parâmetro? Essas propriedades estão baseadas na distribuição de probabilidades do estimador, chamada de distribuição amostral.

2 Distribuições Amostrais

2.1 Introdução

Na seção anterior, vimos que as estatísticas e portanto os estimadores são variáveis aleatórias. A distribuição de probabilidades de uma estatística é conhecida como distribuição amostral e seu desvio-padrão é referido como erro padrão.

Uma forma de obter a distribuição amostral de um estimador é pensarmos em todas as amostras possíveis de tamanho n que podem ser retiradas da população, usando por exemplo, amostragem aleatória simples com reposição.

Como ilustração, consideremos uma população formada por 5 alunos. Temos informação completa sobre a idade e sexo dos alunos na Tabela 1.

Identificação dos alunos	Idade (em anos completos)	Sexo (f=feminino; m=masculino)
1	22	f
2	19	f
3	19	m
4	20	f
5	22	m

Tabela 1: População de alunos

Esta população de alunos tem, em média, $\mu = 20,4$ anos com um desvio-padrão $\sigma = 1,4$ anos e 40% dos alunos são homens, ou seja, a proporção de homens é $P = 0,40$. A idade média, o desvio-padrão de idade e a proporção de homens descrevem a população, portanto são parâmetros.

Na Tabela 2 estão relacionadas todas as amostras possíveis de tamanho dois, que podem ser selecionadas da população de alunos, por amostragem aleatória simples com reposição. Para cada amostra i, podemos calcular $\bar{x}_i$, uma estimativa para a idade média e $\hat{p}_i$, uma estimativa para a proporção de homens.

Temos estimativas diferentes para μ e P, e é o que ocorre quando tiramos várias amostras de uma população.

Por exemplo, a amostra 4, formada pelos alunos 1 e 4, apresenta uma superestimativa para a idade média, $\bar{x}_4 = 21$ anos, e substima a proporção de homens, $\hat{p}_4 = 0,0$. Já na amostra 7, em que foram selecionados os alunos 2 e 3, a idade média é subestimada em $\bar{x}_7 = 19$ anos, e a proporção de homens é superestimada em

Amostra selecionada	Dados amostrais	i	Média amostral $\bar{x}_i$	Proporção amostral $\hat{p}_i$
1 e 1	22 f, 22 f	1	22,0	0,0
1 e 2	22 f, 19 f	2	20,5	0,0
1 e 3	22 f, 19 m	3	20,5	0,5
1 e 4	22 f, 20 f	4	21,0	0,0
1 e 5	22 f, 22 m	5	22,0	0,5
2 e 2	19 f, 19 f	6	19,0	0,0
2 e 3	19 f, 19 m	7	19,0	0,5
2 e 4	19 f, 20 f	8	19,5	0,0
2 e 5	19 f, 22 m	9	20,5	0,5
3 e 3	19 m, 19 m	10	19,0	1,0
3 e 4	19 m, 20 f	11	19,5	0,5
3 e 5	19 m, 22 m	12	20,5	1,0
4 e 4	20 f, 20 f	13	20,0	0,0
4 e 5	20 f, 22 m	14	21,0	0,5
5 e 5	22 m, 22 m	15	22,0	1,0

Tabela 2: Todas as amostras possíveis de tamanho 2 com reposição, da população de alunos

$\hat{p}_7 = 0,5$. Nenhuma das estimativas coincide com o parâmetro. Nesta ilustração, estamos medindo o erro das estimativas, pois o valor do parâmetro é conhecido, situação que não acontece nos problemas reais de estimação.

Como fizemos em estatística descritiva, podemos resumir as informações da Tabela 2 calculando a média e o desvio-padrão das estimativas.

Para as estimativas de μ:

$$\frac{\sum_{i=1}^{15} \bar{x}_i}{15} = \frac{306}{15} = 20,4$$

$$\sqrt{\frac{\sum_{i=1}^{15} (\bar{x}_i - 20,4)^2}{15}} = \sqrt{\frac{16,1}{15}} = 1,04$$

Para as estimativas de P:

$$\frac{\sum_{i=1}^{15} \hat{p}_i}{15} = \frac{6}{15} = 0,4$$

$$\sqrt{\frac{\sum_{i=1}^{15} (\hat{p}_i - 0,4)^2}{15}} = \sqrt{\frac{2,1}{15}} = 0,37$$

Não foi por acaso que as médias das estimativas coincidiram com os valores dos correspondentes parâmetros. Podemos demonstrar que a média dessas estimativas é igual ao parâmetro que está sendo estimado. E podemos demonstrar que o

desvio-padrão das estimativas de μ é

$$\sigma/\sqrt{n}$$

e que o das estimativas de P é

$$\sqrt{P(1-P)/n}$$

Isto quer dizer que quanto maior o tamanho da amostra, menos dispersas serão as estimativas.

Quando obtemos todas as amostras de tamanho dois da nossa população de 5 alunos encontramos estimativas diferentes para o mesmo parâmetro, porem, em média, são iguais ao parâmetro; e tendem a ser mais homogêneas com o aumento da amostra. Este resultado é surpreendente, e junto com os outros achados desta seção, tornam possível o uso de uma amostra para estimar a população.

Se esta fosse a única forma de obter a distribuição amostral, o processo de inferência ficaria inviável para populações reais. Felizmente, não é necessário obter todas as possíveis amostras de tamanho n para construir a distribuição amostral. A teoria de probabilidade fornece um teorema, que é um dos resultados mais importantes da Estatística, pois encontra uma aproximação para a distribuição amostral de $\bar{X}$, sem necessidade de se conhecer muito sobre a população em estudo.

2.2 Teorema Central do Limite (TCL)

Seja uma variável aleatória contínua X. Suponhamos que na população, X tem uma distribuição com média μ e desvio-padrão σ. Uma amostra aleatória de tamanho n é selecionada da população. Quando n é grande o suficiente (em geral, $n \geq 30$), a média amostral, $\bar{X}$, tem distribuição aproximadamente normal com média μ e desvio-padrão $\sigma/\sqrt{n}$.

Se σ é desconhecido, pode ser substituído por sua estimativa s, o desvio-padrão amostral, e mesmo assim, o resultado do teorema não muda muito. Podemos então dizer que

$$\frac{\bar{X}-\mu}{s/\sqrt{n}}$$

tem aproximadamente distribuição normal padrão.

Porque é importante saber a distribuição de $\bar{X}$? Para avaliar a qualidade dos estimadores (Seção 3) e para associar um erro máximo para as estimativas.

Podemos considerar que $\bar{X}$ é um bom estimador de μ quando o erro amostral, $\bar{X}-\mu$, for pequeno. Mas como μ é desconhecido não é possível mensurar o erro amostral.

Outra idéia, é calcular a probabilidade do erro amostral ser no máximo igual a E (para mais ou para menos). Isto é, usando a distribuição de $\bar{X}$, podemos calcular

$$P\left(-E \leq \bar{X}-\mu \leq E\right)$$

Esperamos que essa probabilidade seja bastante alta, perto de 100%.

Como ilustração do uso de distribuições amostrais, consideremos uma amostra aleatória de $n = 100$ trabalhadores imigrantes, com a informação sobre o salário mensal. O salário médio mensal da amostra é $\bar{x} =$ R\$1500,00, uma estimativa de μ, o salário médio mensal de todos os trabalhadores imigrantes. E o desvio-padrão amostral de salário é $s =$ R\$1200,00. Os responsáveis pela pesquisa consideram que uma boa estimativa para o salário médio mensal deve ter no máximo um erro $E =$ R\$200,00.
Vamos calcular a probabilidade do erro máximo ser de 200 reais,

$$P\left(-200 \leq \bar{X} - \mu \leq 200\right)$$

Para tanto, usaremos o TCL, já que $n = 100$.
Dividindo todos os termos dentro da probabilidade por

$$\frac{s}{\sqrt{n}} = \frac{1200}{\sqrt{100}} = 120$$

teremos

$$P\left(-200 \leq \bar{X} - \mu \leq 200\right) = P\left(-1,67 \leq \frac{\bar{X} - \mu}{120} \leq 1,67\right)$$

Pelo TCL,

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

tem aproximadamente distribuição normal padrão. Seja Z uma variável aleatória com distribuição normal padrão. Podemos então escrever

$$P\left(-1,67 \leq \frac{\bar{X} - \mu}{120} \leq 1,67\right) \cong P\left(-1,67 \leq Z \leq 1,67\right) = 0,905$$

em que a segunda probabilidade pode ser obtida da tabela de distribuição normal padrão.
Temos então que

$$P\left(-200 \leq \bar{X} - \mu \leq 200\right) \cong 0,905$$

ou seja, a probabilidade do erro amostral ser de no máximo R\$200,00 é aproximadamente igual a 90,5%. O valor 90,5% é conhecido como nível de confiança. Em outras palavras, não podemos garantir que o erro máximo não será ultrapassado, mas temos $90,5\%$ de confiança que ele não será maior que R\$200,00.

Para aumentar o nível de confiança, é preciso aumentar n, o tamanho da amostra; ou diminuir E, o erro máximo. O processo de encontrar n, fixados o nível de confiança e o erro máximo, é um problema de cálculo de tamanho de amostra, apresentado na Seção 9. O processo de encontrar E, fixados o tamanho de amostra e o nível de confiança, é um problema de estimação. Neste tipo de problema, fazemos o cálculo da probabilidade ao contrário: temos o valor 0,905 e queremos encontrar 1,67. Os detalhes desse processo inverso são vistos na Seção 4.

2.3 TCL para Proporção

A proporção populacional P pode ser vista como uma média. Suponhamos uma população com N elementos. Definamos a variável $X_i = 1$, se o elemento i da população pertence a categoria de interesse, e $X_i = 0$, se não pertence a categoria. Podemos calcular a proporção populacional por

$$P = \frac{\sum_{i=1}^{N} X_i}{N}$$

Então, o cálculo de P é análogo ao de uma média e portanto, é possível aplicar TCL para a proporção amostral.

Seja uma amostra aleatória de tamanho n, selecionada de uma população com proporção populacional igual a P. Quando n é grande o suficiente (em geral, $n \geq 30$, se valor de P não for muito próximo de 0 ou 1), então $\hat{P}$, a proporção amostral, tem aproximadamente uma distribuição normal com média P e desvio-padrão

$$\sqrt{P(1-P)/n}$$

estimado por

$$\sqrt{\hat{p}(1-\hat{p})/n}$$

Assim,

$$\frac{\hat{P} - P}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

tem aproximadamente distribuição normal padrão.

Após entender os vários aspectos básicos da teoria de probabilidade com relação aos estimadores de parâmetros, nós estamos prontos para aventurar na inferência estatística.

Ao inferir partindo de um conjunto limitado de dados (amostra) para o conjunto inteiro de dados (população), estamos lidando com a variabilidade do acaso, portanto erros amostrais são inevitáveis. Já vimos que nunca teremos certeza de que o resultado amostral será o mesmo do correspondente parâmetro, pois não podemos medir o erro amostral, mas podemos calcular a probabilidade do erro máximo. Uma outra forma de avaliar a qualidade dos estimadores é pelas suas propriedades.

3 Propriedades dos Estimadores

Podemos propor vários estimadores para um determinado parâmetro. Para estimar, por exemplo, a média populacional μ da variável X, nós poderíamos usar a média amostral $\bar{X}$, a mediana amostral, ou a primeira observação X_1, entre outras possibilidades. Alguns estimadores em potencial não tem sentido como X_1, que considera a primeira observação como estimador de μ e despreza toda a informação proveniente das outras observações na amostra. Pode ser natural usar a estatística análoga para estimar o parâmetro, ou seja, usar a média amostral para estimar μ,

mas nem sempre o análogo é o melhor estimador. Basicamente, um bom estimador tem uma distribuição amostral com (a) média igual ao parâmetro sendo estimado e (b) erro padrão pequeno.

- Vício

  Um estimador é não viciado se sua distribuição amostral está centrada ao redor do parâmetro, de forma que a sua média é o parâmetro. Pelo TCL, a média amostral é um estimador não viciado da média populacional. Por outro lado, um estimador viciado, em média, tende a subestimar ou sobrestimar o parâmetro. As vezes, pode ser interessante usar estimadores viciados, com vícios que tendem a desaparecer quando o tamanho da amostra aumenta.

- Eficiência

  Uma segunda propriedade interessante para um estimador é ter um pequeno erro padrão, comparado a outros estimadores. Um estimador com essa propriedade é dito ser eficiente.

Um bom estimador de um parâmetro deve ser não viciado e eficiente. Vimos que a média amostral e a proporção amostral são estimadores não viciados para μ e P, respectivamente; e é possível mostrar que são estimadores eficientes. Por outro lado,

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

é um estimador viciado de σ e

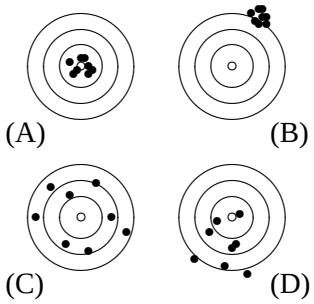
$$\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}$$

é não viciado.

Uma boa maneira de visualizar as propriedades dos estimadores, é fazer uma analogia com o jogo de dardos. Na Figura 1 estão esquematizados o desempenho de 4 jogadores, cada um com 8 dardos. Os dardos são as amostras e os jogadores representam 4 tipos de estimadores. O jogador da Figura 1A representa um bom estimador, pois os dardos estão em torno do alvo (não viciado) e bem concentrados (eficiente). Nas Figuras 1B a 1D os jogadores não tem um desempenho tão bom. Na Figura 1B está representado o estimador mais eficiente, comparando com os outros estimadores, mas tem vícios. Já o estimador caracterizado na Figura 1C não tem vícios porém, não é eficiente. O jogador da Figura 1D representa o pior dos 4 estimadores: é viciado e não pode ser considerado eficiente.

Até o momento apresentamos uma estimativa do parâmetro baseando-nos em um único valor, referido como estimação pontual. Por exemplo, se a proporção de mulheres com osteoporose em uma comunidade for estimada em 45%, essa estimativa é pontual pois se baseia em um único valor numérico. Muitas vezes,

Figura 1: Analogia entre as propriedades dos estimadores e o jogo de dardos



entretanto, queremos considerar, conjuntamente, o estimador e a sua variabilidade. A forma usual de incorporar esta informação é por meio do chamado intervalo de confiança.

Com uma amostra disponível, nós podemos usar as propriedades da distribuição amostral do estimador para formar um intervalo de confiança, isto é, um intervalo de valores que deve conter o verdadeiro valor do parâmetro com uma probabilidade pré-determinada, referida por nível de confiança. Neste tipo de estimação, nós acreditamos, com um certo nível de confiança, que o intervalo contém o valor do parâmetro. Um exemplo, seria dizer que a proporção estimada de mulheres com osteoporose está entre 40% e 50% com um nível de 95% de confiança.

4 Estimação de uma Média Populacional (μ)

Consideremos uma população em que há interesse em estimar, μ, a média de X, uma variável aleatória contínua. Suponhamos que uma amostra aleatória de tamanho n foi selecionada da população para obter uma estimativa de μ.

O estimador $\bar{X}$ é aquele que apresenta as melhores propriedades para estimar média populacional, μ.

- Intervalo de Confiança

Suponhamos que n seja suficientemente grande para a aplicação do TCL. Então a quantidade

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

tem aproximadamente distribuição normal padrão.

Seja Z uma variável aleatória com distribuição normal padrão. Como representado na Figura 2, podemos encontrar o valor de z, tal que

$$P\left(-z \leq Z \leq z\right) = 1 - \alpha$$

pela tabela da distribuição normal padrão.

Figura 2: Densidade de Z e o quantil z

Temos uma quantidade que é aproximadamente normal e temos Z que é normal. Podemos então escrever

$$P\left(-z \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq z\right) \cong 1-\alpha$$

Isolando μ, ao passar os demais elementos para o outro lado das duas desigualdades dentro da probabilidade, temos que

$$P\left(\bar{X}-z\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X}+z\frac{s}{\sqrt{n}}\right) \cong 1-\alpha \tag{1}$$

Em 1, a interpretação da probabilidade é diferente. Entre as desigualdades está o parâmetro μ, e não são calculadas probabilidades de parâmetros, pois são valores fixos. As partes aleatórias estão no outro lado de cada uma das desigualdades. A melhor interpretação para 1 é que com $(1-\alpha)$ de confiança, o intervalo

$$\left(\bar{X}-z\frac{s}{\sqrt{n}}; \bar{X}+z\frac{s}{\sqrt{n}}\right)$$

contém o parâmetro.

Com os dados amostrais, calculamos o intervalo, usando a estimativa $\bar{x}$. Então, um intervalo de confiança de $(1-\alpha)$ é dado por

$$\left(\bar{x}-z\frac{s}{\sqrt{n}}; \bar{x}+z\frac{s}{\sqrt{n}}\right)$$

As vezes, é interessante avaliar o erro máximo da estimativa, E, ao invés de construir o intervalo de confiança. Para um nível de confiança $(1-\alpha)$

$$E = z\frac{s}{\sqrt{n}}$$

em que z pode ser obtido da tabela da normal padrão, em função do nível de confiança desejado.

Exemplo 1 *Um centro de ortodontia deseja conhecer a estimativa do tempo médio que um membro da equipe gasta para atender a cada paciente. Suponha que uma amostra de 38 especialistas revelou que a média foi de 45 minutos com um desvio-padrão de 6 minutos. Determine um intervalo de 99% de confiança para o parâmetro.*

Desejamos uma estimativa para o parâmetro desconhecido μ, o tempo médio que um membro da equipe gasta para atender um paciente. Temos que $n = 38$, $\bar{x} = 45$ e $s = 6$. Para um nível de confiança de 99%, o valor tabelado para z é 2,58. Portanto, o intervalo de 99% de confiança para μ é

$$\left(45 - 2,58\frac{6}{\sqrt{38}}; 45 + 2,58\frac{6}{\sqrt{38}}\right)$$
$$(42,49; 47,51)$$

Podemos dizer então que o intervalo entre 42,49 e 47,51 contém o tempo médio gasto por um membro da equipe para atender a cada paciente, com 99% de confiança.

5 Estimação de μ em Amostras Pequenas

Quando dispomos de uma amostra pequena $(n < 30)$, não temos a garantia da aplicação do TCL, portanto a distribuição amostral da média pode ou não estar próxima da distribuição normal.

A teoria de probabilidades mostra que, mesmo assim, é possível construir estimativas intervalares para média populacional, μ, utilizando uma certa distribuição, denominada t de Student. Esta distribuição tem forma parecida com a da normal padrão, com caudas um pouco mais pesadas (ver Figura 3), ou seja, a dispersão da distribuição t de Student é maior.

Figura 3: Densidade de T e Z

Esta dispersão varia com o tamanho da amostra, sendo bastante dispersa para amostras pequenas, mas se aproximando da normal padrão para amostras grandes. A distribuição t de Student tem apenas um parâmetro, denominado graus de liberdade, gl. No caso da estimação de uma média, $gl = n - 1$.

Pela teoria de probabilidades, podemos demonstrar que se uma variável aleatória, X, tem distribuição normal com média μ, então

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

tem distribuição t de Student com $(n - 1)$ gl; e S é o estimador do desvio-padrão de X, σ.

O valor de t, tal que

$$P\left(-t \leq T \leq t\right) = 1 - \alpha \tag{2}$$

pode ser encontrado usando a tabela da distribuição t de Student.

Reescrevendo 2

$$P\left(-t \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t\right) = 1 - \alpha$$

Isolando μ das desigualdades dentro da probabilidade, resulta em

$$P\left(\bar{X} - t\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Usando as estimativas $\bar{x}$ e s, um intervalo de confiança para μ com base em uma amostra pequena é dado por

$$\left(\bar{x} - t\frac{s}{\sqrt{n}}; \bar{x} + t\frac{s}{\sqrt{n}}\right)$$

e o erro máximo de estimação é

$$E = t\frac{\sigma}{\sqrt{n}}$$

em que t depende do nível de confiança desejado.

Para que a distribuição t de Student seja aplicável, a distribuição da população deve ser normal, porém, se tem apenas uma moda e é basicamente simétrica, obtemos em geral, bons resultados, como intervalos de confiança precisos. Se há forte evidência de que a população tem distribuição bastante assimétrica, então uma alternativa é utilizar métodos não-paramétricos, descritos no Capítulo de Inferência.

Exemplo 2 *Uma rede de lanchonetes deseja estimar a quantia média que cada cliente gasta por lanche. Foram coletadas dados de uma amostra de 22 clientes que revelou um uma quantia média de R$ 15 com um desvio-padrão de 5. Construir um intervalo de confiança de 95% para a média populacional.*
O objetivo é estimar o parâmetro μ, a quantia média que cada cliente gasta por lanche de trabalho, em todas as lanchonetes da rede. Podemos obter uma estimativa para o erro padrão da média dado por

$$\frac{5}{\sqrt{22}} = 1,066$$

Usando nível de 95% de confiança e graus de liberdade $gl = 21$ (pois, $n = 22$ e $gl = n - 1$), obtemos na tabela da distribuição t de Student o valor $t = 2,08$, e assim podemos calcular o erro máximo da estimativa

$$E = t\frac{s}{\sqrt{n}} = 2,08 \cdot 1,066 = 2,217$$

Então, temos o seguinte intervalo de 95% de confiança para o parâmetro μ:

$$\left(\bar{x} - t\frac{s}{\sqrt{n}}; \bar{x} + t\frac{s}{\sqrt{n}}\right)$$

$$(15 - 2,217; 15 + 2,217)$$

$$(12,783; 17,217)$$

O intervalo de 12,783 a 17,217 contém a quantia média que cada cliente gasta por lanche, com 95% de confiança.

Pode parecer um pouco estranho que, com uma população distribuída normalmente, venhamos eventualmente a utilizar a distribuição t de Student para achar os valores associados ao nível de confiança; mas quando σ não é conhecido, a utilização de s de uma amostra pequena incorpora outra fonte de erro. Para manter o grau desejado de confiança compensamos a variabilidade adicional ampliando o intervalo de confiança por um processo que substitui o valor z por um valor maior, t. Para ilustrar esta idéia considere para o Exemplo 2 um intervalo de confiança de 95%, utilizando o valor z de uma distribuição normal. Este intervalo apresenta uma amplitude menor.

$$(15 - 2,089; 15 + 2,089)$$

$$(12,911; 17,089)$$

6 Estimação de uma Proporção Populacional (P)

Nesta seção as variáveis são referentes a contagens, como o número de fumantes, número de unidades defeituosas em uma linha de produção, e assim por diante. Inicialmente, abordamos a estimativa pontual do parâmetro P, a proporção populacional, e, em seguida, construiremos estimativas intervalares.

Consideremos o caso em que o parâmetro a ser estimado é a proporção P de indivíduos em uma população, que apresentam uma certa característica. Retira-se da população uma amostra de tamanho n. X será o número de elementos da amostra que apresentam a característica em estudo. Um estimador da proporção populacional P será a proporção amostral $\hat{P}$:

$$\hat{P} = \frac{X}{n}$$

- Intervalo de Confiança

Para estimar, por intervalo, o parâmetro P, a partir de $\hat{P}$, podemos seguir os mesmos princípios da estimação da média populacional.

Suponhamos que n seja suficientemente grande para aplicar o TCL para proporção. Temos que

$$P\left(-z \leq Z \leq z\right) = 1 - \alpha \Rightarrow P\left(-z \leq \frac{\hat{P} - P}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq z\right) \cong 1 - \alpha$$

em que o valor de z é encontrado na tabela da distribuição normal padrão.

Isolando P nas desigualdades dentro da segunda probabilidade, resulta em

$$P\left(\hat{P} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq P \leq \hat{P} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha \qquad (3)$$

A probabilidade em 3 pode ser interpretada como, a probabilidade do intervalo

$$\left(\hat{P} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{P} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

conter o parâmetro, com $(1-\alpha)$ de confiança.

Com os dados da amostra, um intervalo de confiança otimista é

$$\left(\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

e um intervalo de confiança conservativo é

$$\left(\hat{p} - z\sqrt{\frac{1}{4n}}; \hat{p} + z\sqrt{\frac{1}{4n}}\right)$$

em que z é encontrado na tabela da distribuição normal padrão, de acordo com o nível de confiança, $(1-\alpha)$.

A abordagem conservativa substitui o produto $P(1-P)$ por $1/4$. Como indicado na Figura 4, o produto $P(1-P)$ é no máximo igual a $1/4$. De modo que, ao substituir por $1/4$, obtemos o intervalo de confiança mais largo.

Figura 4: Máximo de $P(1-P)$

Os erros máximos de estimação, E, para um nível de confiança $(1-\alpha)$ na abordagem otimista e conservativa são, respectivamente, dados por

$$E = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

e

$$E = z\sqrt{\frac{1}{4n}}$$

Exemplo 3 *Um especialista em educação pretende avaliar a aceitação de um projeto educacional numa cidade. Depois de apresentá-lo às escolas do município, os responsáveis por sua execução desejam avaliar o valor aproximado do parâmetro P, a proporção de diretores favoráveis ao projeto, dentre as escolas do município. Para estimar este parâmetro, o especialista planeja observar uma amostra aleatória simples de $n = 600$ escolas. Por exemplo, se na amostra 420 são favoráveis, temos a seguinte estimativa pontual para o parâmetro P:*

$$\hat{p} = \frac{420}{600}$$

$$\hat{p} = 0,70$$

Usando um nível de 95% de confiança temos o seguinte intervalo otimista:

$$\left(0,7-1,96\sqrt{\frac{0,70\cdot 0,30}{600}};0,70+1,96\sqrt{\frac{0,70\cdot 30}{600}}\right)$$

$$(0,663;0,737)$$

Ou seja, o intervalo de 66,3% a 73,7% contém, com 95% de confiança, a porcentagem de favoráveis ao projeto, dentre todas as escolas municipais.

7 Estimação da Diferença entre Duas Médias Populacionais μ_1 e μ_2

Em muitas situações há necessidade de comparar duas populações diferentes. Como exemplos, podemos ter interesse em saber: se idosos que praticam exercícios físicos diariamente apresentam nível de colesterol menor do que idosos, com as mesmas condições, mas que não praticam exercícios físicos diariamente; se um tipo de equipamente eletrônico tem maior durabilidade do que outro; e assim por diante. A seguir vamos utilizar métodos para construir um intervalo de confiança para a diferença entre duas médias, com duas amostras independentes. Para tanto uma amostra aleatória é selecionada independentemente de cada população.

Seja X a variável aleatória a ser comparada. Suponhamos que X tenha média μ_1 e desvio-padrão σ_1 na População 1 e tenha média μ_2 e desvio-padrão σ_2 na População 2. Suponhamos também que as amostras tenham tamanhos n_1 e n_2 para as Populações 1 e 2, respectivamente. Se n_1 e $n_2 \geq 30$, pelo TCL, pode ser mostrado que $\bar{X}_1 - \bar{X}_2$ tem distribuição aproximadamente normal com média $\mu_1 - \mu_2$ e desvio-padrão

$$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

Quando os parâmetros, σ_1 e σ_2, são desconhecidos podem ser substituídos por s_1 e s_2, os desvios-padrão amostrais.

Usando o mesmo processo para a construção de intervalo de confiança para uma média, o intervalo de confiança para $(\mu_1 - \mu_2)$ será dado por:

$$\left((\bar{x}_1-\bar{x}_2)-z\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}};(\bar{x}_1-\bar{x}_2)+z\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)$$

Exemplo 4 *Com o objetivo de comparar dois métodos de redução de gordura localizada nas coxas, foram criados dois grupos, cada um com 30 pessoas que apresentam as mesmas condições, recebendo um tipo de tratamento. Antes e depois de um período de 60 dias de utilização do aparelho foram anotados as perdas em mm. Obtendo-se:*

$$\bar{x}_1 = 21,3\,;\;\; s_1 = 2,6$$

$$\bar{x}_2 = 13,4\,;\;\; s_2 = 1,9$$

Construir o intervalo de 95% de confiança para a diferença de médias.
A diferença é

$$\bar{x}_1 - \bar{x}_2 = 7,9$$

$$\left(7,9 - 1,96\sqrt{\frac{2,6^2}{30} + \frac{1,9^2}{30}}; 7,9 + 1,96\sqrt{\frac{2,6^2}{30} + \frac{1,9^2}{30}}\right)$$

$$(6,748; 9,0527)$$

é o intervalo de confiança de 95% de confiança para a diferença entre as reduções médias dos dois métodos.

Quando o zero está no intervalo de confiança significa que não há diferença entre as duas médias populacionais. No Exemplo 4, com base no intervalo de confiança de 95%, podemos concluir que há diferença entre as médias de redução de gordura das coxas entre os dois métodos utilizados.

8 Estimação de $\mu_1 - \mu_2$ em Amostras Pequenas

Se n_1 ou n_2 é menor que 30, fórmulas alternativas devem ser usadas para os intervalos de confiança. A teoria usada na Seção 5 pode ser estendida para o caso de duas amostras. Assumimos normalidade para as distribuições populacionais, como no caso de uma amostra. Além disso, devemos assumir aqui que $\sigma_1 = \sigma_2$, isso quer dizer que as duas populações tem o mesmo desvio-padrão, digamos, σ. A variância populacional, σ^2 que pode ser estimada por

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

O intervalo de confiança para $\mu_1 - \mu_2$ será dado por:

$$\left((\bar{x}_1 - \bar{x}_2) - t\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}; (\bar{x}_1 - \bar{x}_2) + t\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

Exemplo 5 *O tempo para realizar uma tarefa, em segundos, foi anotado para 10 homens e 11 mulheres, igualmente treinados. As médias e variâncias obtidas foram:*

Determine um intervalo de confiança de 99% para a diferença entre os tempos médios de homens e mulheres.

Homem	Mulher
$\bar{x}_1 = 45,33$	$\bar{x}_2 = 43,54$
$s_1^2 = 1,54$	$s_2^2 = 2,96$

Primeiramente, vamos calcular s_p

$$s_p^2 = \sqrt{\frac{9 \cdot 1,54 + 10 \cdot 2,96}{19}} = 2,29$$

Na tabela da distribuição de t de Student com 19 gl encontramos que $t = 2,78$*, para 99% de confiança.*
$\bar{x}_1 - \bar{x}_2 = 1,79$

O intervalo de confiança para $(\mu_1 - \mu_2)$ *será dado por:*

$$\left(1,79 - 2,78\sqrt{2,29\left(\frac{1}{10} + \frac{1}{11}\right)}; 1,79 + 2,78\sqrt{2,29\left(\frac{1}{10} + \frac{1}{11}\right)}\right)$$

$$(-0,047; 3,63)$$

Com base em um intervalo de confiança de 99%, não existe diferença entre os tempos médios de homens e mulheres.

9 Determinação do tamanho da amostra (n)

Supondo que há condições para aplicação do TCL, as fórmulas para o cálculo de n são derivadas, fixando o erro máximo, E e o nível de confiança $(1 - \alpha)$. A determinação de n também depende do plano amostral adotado e do parâmetro sendo estimado.

No caso da amostragem aleatória simples, a fórmula de n para a estimação de μ é encontrada isolando n em

$$E = z\frac{\sigma}{\sqrt{n}}$$

Portanto,

$$n = \left(\frac{z\sigma}{E}\right)^2$$

em que σ deve ser previamente estimado e z é obtido conforme o nível de confiança.

Sugestões para estimação prévia de σ:

1. usar estimativas de σ, de um estudo similar feito anteriormente ou de uma amostra piloto;

2. Em muitas situações, podemos considerar que $\sigma \approx \frac{\text{amplitude}}{4}$. O argumento teórico para o uso desta aproximação está baseado na propriedade da distribuição normal com média μ e desvio-padrão σ, de que a área entre $\mu - 2\sigma$ e $\mu + 2\sigma$ é igual a 95,5%.

Exemplo 6 *Qual é o tamanho de amostra necessário para estimar a renda média mensal das famílias de uma pequena comunidade, com um erro máximo de R$100,00 com 95% de confiança, usando amostragem aleatória simples? Sabe-se que a renda mensal famliar está entre R$50,00 e R$1000,00. Temos que* $E = 200$ *e para um nível de confiança igual a 95% temos que* $z = 1,96$*. Com a informação de que a renda varia entre 50 e 10000, uma aproximação para* $sigma$ *é*

$$\sigma \approx \frac{\text{amplitude}}{4} = \frac{1000 - 50}{4} = 237,5$$

Assim,

$$n = \left(\frac{z\sigma}{E}\right)^2 = \left(\frac{1,96 \cdot 237,5}{100}\right)^2 = 21,67 \approx 22$$

Portanto, cerca de 22 famílias devem ser entrevistadas.

Para a estimação de P, a fórmula para determinar n, usando amostragem aleatória simples, é encontrada isolando n em

$$E = z\sqrt{\frac{P(1-P)}{n}}$$

Portanto,

$$n = z^2\frac{P(1-P)}{E^2}$$

em que P deve ser previamente estimado e z é obtido conforme o nível de confiança.

Sugestões para estimação prévia de P:

1. usar estimativas de P de um estudo similar feito anteriormente ou de uma amostra piloto;

2. substituir o produto $P(1-P)$ por 0,25. Notamos que ao substituir por 0,25, o tamanho da amostra pode ser maior que o necessário.

Exemplo 7 *Líderes estudantis de uma faculdade querem conduzir uma pesquisa para determinar a proporção* P *de estudantes a favor de uma mudança no horário de aulas. Como é impossível entrevistar todos os 2000 estudantes em um tempo razoável, decide-se fazer uma amostragem aleatória simples dos estudantes:*

a) Determinar o tamanho de amostra (número de estudantes a serem entrevistados) necessário para estimar P *com um erro máximo de 0,05 e nível de confiança de 95%. Assumir que não há nenhuma informação a priori disponível para estimar* P*; Temos que* $E = 200$ *e que* $z = 1,96$*. Como não há informação a priori sobre* P*,*

$$n = z^2\frac{P(1-P)}{E^2} = 1,96^2\frac{0,25}{0,05^2} = 384,16 \approx 385$$

Para estimar o proporção de estudantes favoráveis a mudança de horário, é necessária uma amostra de 385 estudantes.

b) Os líderes estudantis também querem estimar a proporção de P' de estudantes que sentem que a representação estudantil atende adequadamente as suas necessidades. Com um erro máximo de 7% e nível de confiança de 95%, determinar o tamanho de amostra para estimar P'. Utilizar a informação de uma pesquisa similar conduzida anos atrás, quando 60% dos estudantes acreditavam que estavam bem representados;

$$n = z^2 \frac{P'(1-P')}{E^2} = 1,96^2 \frac{0,60(1-0,60)}{0,07^2} = 188,16 \approx 189$$

Para estimar o proporção de estudantes que se consideram bem representados, é necessária uma amostra de 189 estudantes.

c) Qual o tamanho de amostra adequado para atingir ambos os objetivos da pesquisa?

Para atingir ambos os objetivos da pesquisa, devemos considerar a maior amostra, que é a de 385 estudantes.

Quando N (tamanho da população) é conhecido, o valor de n para estimar μ e P pode ser corrigido (n^*):

$$n^* = \frac{Nn}{N+n}$$

Notamos que se N é muito maior que n, então n^* é aproximadamente n.

Exemplo 8 *Determinar o tamanho de amostra necessário para estimar o volume médio de vendas de carros novos nacionais entre as concessionárias, fixando um nível de confiança de 99% para um erro de estimação igual a 1 automóvel. É conhecido que existem 200 concessionárias na região em estudo. Em uma pesquisa similar feita a 5 anos atrás, o desvio-padrão amostral foi igual a 2,8. Supor que foi feita uma amostragem aleatória simples.*

Temos que $E = 1$ e para um nível de confiança igual a 99% temos que $z = 2,58$. Usaremos a estimativa a priori para σ, substituindo-o na fórmula por 2,8. Assim,

$$n = \left(\frac{z\sigma}{E}\right)^2 = \left(\frac{2,58 \cdot 2,8}{1}\right)^2 = 52,19 \approx 53$$

Com a informação de que há $N = 200$, podemos corrigir o valor de n

$$n^* = \frac{Nn}{N+n} = \frac{200 \cdot 53}{200+53} = \frac{10600}{253} = 41,90 \approx 42$$

Portanto, é necessário selecionar 42 concessionárias de automóveis.