C

Fundamentos de Estatística Matemática

C.1 POPULAÇÕES, PARÂMETROS E AMOSTRAGEM ALEATÓRIA

A inferência estatística envolve o conhecimento de dados sobre uma população, dada a disponibilidade de uma amostra dessa população. Por **população**, entendemos qualquer grupo de tópicos bem definido, que poderia ser de indivíduos, empresas, cidades, ou muitas outras possibilidades. Por "conhecimento" podemos entender várias coisas, que de um modo geral dividimos nas categorias de *estimação* e *testes de hipóteses*.

Alguns exemplos podem ajudar a compreender esses termos. Na população de todos os adultos trabalhadores nos Estados Unidos, os economistas especializados na área de trabalho estão interessados em estudar o retorno da educação, indicado pelo aumento percentual médio nos rendimentos, dado mais um ano de educação. Seria impraticável e caro obter informações sobre os rendimentos e graus de educação da totalidade da população trabalhadora nos Estados Unidos, mas podemos obter dados de um subconjunto da população. Usando os dados coletados, um economista poderá informar que sua melhor estimativa do retorno de mais um ano de educação é de 7,5%. Esse é um exemplo da *estimativa por ponto*. Ou, poderá descrever uma faixa, tal como "o retorno da educação está entre 5,6% e 9,4%". Esse é um exemplo de uma *estimativa por intervalo*.

Um economista especializado em urbanismo pode querer saber se assistir a programas de televisão sobre prevenção de crimes em determinada região está associado a índices de criminalidade mais baixos na vizinhança dessa região. Após comparar os índices de criminalidade da redondeza com e sem tais programas em uma amostra da população, ele poderá chegar a uma de duas conclusões: os programas de TV sobre prevenção de crimes realmente afetam a criminalidade, ou não. Esse exemplo situa-se na rubrica dos testes de hipóteses.

O primeiro passo na inferência estatística é identificar a população de interesse. Isso pode parecer óbvio, mas é importante ser bastante específico. Logo que tenhamos identificado a população, poderemos especificar um modelo para a relação populacional de interesse. Tais modelos envolvem distribuições de probabilidade ou características de distribuições de probabilidade, e elas dependem de parâmetros desconhecidos. Parâmetros são simplesmente constantes que determinam as direções e intensidades da relação entre variáveis. No exemplo anterior da economia do trabalho, o parâmetro de interesse é o retorno da educação na população.

Amostragem

Para revisar a inferência estatística, concentramo-nos no cenário mais simples possível. Seja Y uma variável aleatória representando uma população com uma função de densidade de probabilidade $f(y;\theta)$, que depende do único parâmetro θ. A função de densidade de probabilidade (fdp) de Y é assumida como conhecida, exceto quanto ao valor de θ; valores diferentes de θ implicam diferentes distribuições populacionais, e, portanto, estamos interessados no valor de θ. Se pudermos obter certos tipos de amostras da população, então, poderemos descobrir alguma coisa sobre θ. O esquema de amostragem mais fácil de trabalhar é a amostragem aleatória.

AMOSTRAGEM ALEATÓRIA

Se $Y_1, Y_2, ..., Y_n$ forem variáveis aleatórias independentes com uma função de densidade de probabilidade $f(y;\theta)$ comum, então, $\{Y_1, Y_2, ..., Y_n\}$ é definida como uma **amostra aleatória** a partir de $f(y;\theta)$ [ou uma amostra aleatória a partir da população representada por $f(y;\theta)$].

Quando $\{Y_1, ..., Y_n\}$ é uma amostra aleatória a partir da função de densidade $f(y;\theta)$, também dizemos que as Y_i são amostras *independentes e identicamente distribuídas* (ou i.i.d.) a partir de $f(y;\theta)$. Em alguns casos, não precisaremos especificar em sua totalidade qual é a distribuição comum.

A natureza aleatória de $Y_1, Y_2, ..., Y_n$ na definição de amostragem aleatória reflete o fato que são possíveis muitos resultados diferentes antes da amostragem ter sido efetivamente realizada. Por exemplo, se a renda familiar for obtida de uma amostra de $n = 100$ famílias nos Estados Unidos, as rendas que observaremos em geral diferirão para cada amostra diferente de 100 famílias. Uma vez obtida uma amostra, teremos um conjunto de números, digamos, $\{y_1, y_2, ..., y_n\}$, que constituirá os dados com os quais trabalharemos. Se é ou não apropriado assumir que a amostra é proveniente de um esquema aleatório de amostragem, exige conhecimento sobre o efetivo processo de amostragem.

Amostras aleatórias a partir de uma distribuição de Bernoulli são freqüentemente usadas para ilustrar conceitos estatísticos, e elas também surgem em aplicações empíricas. Se $Y_1, Y_2, ..., Y_n$ forem variáveis aleatórias independentes e cada uma for distribuída como Bernoulli(θ), de forma que $P(Y_i = 1) = 0$ e $P(Y_i = 0) = 1 - \theta$, então, $\{Y_1, Y_2, ..., Y_n\}$ constituirá uma amostra aleatória a partir da distribuição de Bernoulli(θ). Como ilustração, considere o exemplo das reservas da empresa aérea desenvolvido no Apêndice B. Cada Y_i mostra se o passageiro i comparece para embarque; $Y_i = 1$ se o passageiro comparece e $Y_i = 0$, caso contrário. Dessa forma, θ é a probabilidade de uma pessoa, escolhida aleatoriamente na população de todas as pessoas que fizeram reserva, comparecer para o embarque.

Em muitas outras aplicações, as amostras aleatórias podem ser assumidas como retiradas de uma distribuição normal. Se $\{Y_1, ..., Y_n\}$ for uma amostra aleatória a partir de uma população Normal(μ,σ^2), então, a população será caracterizada por dois parâmetros, a média μ e a variância σ^2. O interesse principal geralmente reside em μ, mas σ^2 é de interesse por si mesma, pois fazer inferências sobre μ freqüentemente exige conhecimento de σ^2.

C.2 PROPRIEDADES DOS ESTIMADORES EM AMOSTRAS FINITAS

Nesta seção, estudaremos as chamadas propriedades dos estimadores em amostras finitas. O termo "amostra finita" advém do fato de que as propriedades são válidas para uma amostra de qualquer tamanho, não importando o quanto ela é grande ou pequena. Algumas vezes, elas são chamadas de propriedades de amostras pequenas. Na Seção C.3, trataremos das "propriedades assimptóticas", que estão relacionadas ao comportamento dos estimadores conforme o tamanho da amostra cresce sem limites.

ESTIMADORES E ESTIMATIVAS

Para estudar as propriedades dos estimadores, devemos definir o que entendemos por estimador. Dada uma amostra aleatória $\{Y_1, Y_2, ..., Y_n\}$ retirada de uma distribuição populacional que dependa de um parâmetro desconhecido θ, um **estimador** de θ é uma regra que atribui a cada resultado possível da amostra um valor de θ. A regra é especificada antes de extrair qualquer amostra; em particular, a regra será a mesma independentemente dos dados efetivamente obtidos.

Como um exemplo de um estimador, seja $\{Y_1, ..., Y_n\}$ uma amostra aleatória de uma população com média μ. Um estimador natural de μ é a média da amostra aleatória:

$$\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i. \tag{C.1}$$

$\bar{Y}$ é chamado de **média amostral**, mas, diferentemente do discutido no Apêndice A, no qual definimos a média amostral de um conjunto de números como uma estatística descritiva, $\bar{Y}$ agora é visto como um estimador. Dado qualquer resultado das variáveis aleatórias $Y_1, ..., Y_n$, usamos a mesma regra para estimar μ: simplesmente calculamos suas médias. Para resultados de dados efetivos $\{y_1, ..., y_n\}$, a **estimativa** será simplesmente a média da amostra: $\bar{Y} = (y_1 + y_2 + ... + y_n)/n$.

EXEMPLO C.1

(Taxas de Desemprego nas Cidades)

Suponha que obtemos a seguinte amostra de taxas de desemprego de dez cidades nos Estados Unidos:

Cidade	Taxa de Desemprego
1	5,1
2	6,4
3	9,2
4	4,1
5	7,5
6	8,3
7	2,6
8	3,5
9	5,8
10	7,5

Nossa estimativa da taxa média de desemprego nas cidades dos Estados Unidos será $\bar{y} = 6{,}0$. Cada amostra geralmente resulta em uma estimativa diferente. Porém, a regra para obter a estimativa é a mesma, independente de quais ou quantas cidades aparecem na amostra.

De forma mais geral, um estimador W de um parâmetro θ pode ser expresso como uma fórmula matemática resumida:

$$W = h(Y_1,Y_2, ..., Y_n), \tag{C.2}$$

para alguma função h conhecida das variáveis aleatórias $Y_1, Y_2, ..., Y_n$. Como no caso especial da média amostral, W é uma variável aleatória, porque ela depende da amostra aleatória: se obtivermos diferentes amostras aleatórias da população, o valor de W pode mudar. Quando um conjunto particular de números, digamos $\{y_1, y_2, ..., y_n\}$, é agregado na função h, obtemos uma estimativa de θ, representada por $w = h(y_1, y_2, ..., y_n)$. Algumas vezes W é chamado de estimador por ponto e w de estimativa por ponto, para distingui-los dos estimadores por intervalo e das estimativas por intervalo, aos quais retornaremos na seção C.5.

Para avaliar os procedimentos de estimação, estudamos várias propriedades da distribuição de probabilidade da variável aleatória W. A distribuição de um estimador é muitas vezes chamada de sua **distribuição amostral**, pois essa distribuição descreve a probabilidade de vários resultados de W entre diferentes amostras aleatórias. Como há um número ilimitado de regras para combinar dados para estimar parâmetros, precisamos de algum critério lógico para fazer a escolha entre os estimadores, ou pelo menos para eliminar a consideração de alguns estimadores. Portanto, devemos abandonar o âmbito da estatística descritiva, na qual calculamos coisas como média amostral para simplesmente resumir um acervo de dados. Na estatística matemática, estudamos as distribuições amostrais dos estimadores.

Inexistência de Viés

Em princípio, a totalidade da distribuição amostral de W pode ser obtida, dada a distribuição de probabilidade de Y_i e a função h. Em geral, é mais fácil enfatizar algumas poucas características da distribuição de W ao o avaliarmos como um estimador de θ. A primeira propriedade importante de um estimador envolve seu valor esperado.

ESTIMADOR NÃO-VIESADO

Um estimador W de θ será *não-viesado* se

$$E(W) = \theta, \tag{C.3}$$

para todos os possíveis valores de θ.

Se um estimador for não-viesado, então, sua distribuição de probabilidade terá um valor esperado igual ao parâmetro que ele supostamente estará estimando. A **inexistência de viés** *não* significa que a estimativa que obteremos com qualquer amostra particular será igual a θ, ou mesmo muito próxima de θ. Particularmente, se pudéssemos extrair *indefinidamente* amostras aleatórias de Y da população, calcular uma estimativa a cada vez, e depois calcularmos a média dessas estimativas de todas as amostras aleatórias, obteríamos θ. Esse experimento ideal é abstrato porque, na maior parte das aplicações, temos apenas uma amostra aleatória com que trabalhar.

Para um estimador viesado, definimos seu viés conforme segue.

VIÉS DE UM ESTIMADOR

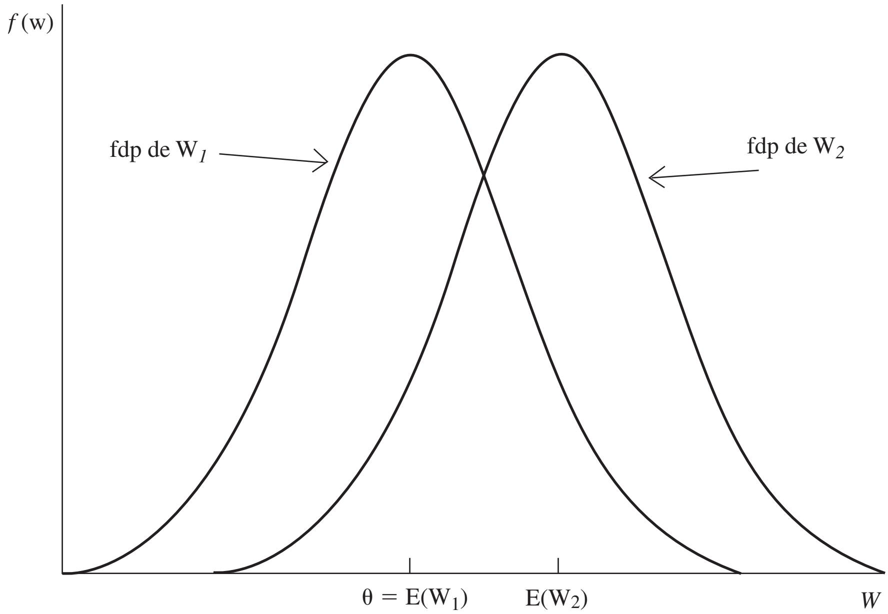
Se W for um estimador de θ, seu **viés** é definido como

$$\text{Viés}(W) \equiv \text{E}(W) - \theta. \tag{C.4}$$

A Figura C.1 mostra dois estimadores; o primeiro não tem viés, e o segundo tem um viés positivo.

Figura C.1

Um estimador sem viés, W_1, e um estimador com viés positivo, W_2.



A inexistência de viés em um estimador e o tamanho de qualquer possível viés dependem da distribuição de Y e da função h. A distribuição de Y geralmente está fora de nosso controle (embora freqüentemente escolhamos um *modelo* para essa distribuição): ela pode ser determinada pela natureza ou por forças sociais. Entretanto, a escolha da regra h é nossa, e se quisermos um estimador não-viesado, então, precisaremos escolher h de maneira apropriada.

É possível mostrar que alguns estimadores podem ser não-viesados de forma bastante genérica. Mostraremos agora que a média amostral $\bar{Y}$ é um estimador não-viesado da média populacional μ, independente da distribuição populacional subjacente. Usamos as propriedades dos valores esperados (E.1 e E.2) das quais tratamos na seção B.3:

$$\mathrm{E}(\bar{Y}) = \mathrm{E}\left((1/n)\sum_{i=1}^{n} Y_i\right) = (1/n)\mathrm{E}\left(\sum_{i=1}^{n} Y_i\right) = (1/n)\left(\sum_{i=1}^{n}\mathrm{E}(Y_i)\right)$$

$$= (1/n)\left(\sum_{i=1}^{n} \mu\right) = (1/n)(n\mu) = \mu.$$

Para os testes de hipóteses, precisaremos estimar a variância σ^2 de uma população com média μ. Definindo $\{Y_1, ..., Y_n\}$ como a amostra aleatória da população com $\mathrm{E}(Y) = \mu$ e $\mathrm{Var}(Y) = \sigma^2$, definimos o estimador como

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \tag{C.5}$$

que normalmente é chamado de **variância amostral**. É possível mostrar que S^2 é um estimador não-viesado de σ^2: $E(S^2) = \sigma^2$. A divisão por $n - 1$, em lugar de n, leva em conta o fato de que a média μ é estimada, em vez de conhecida. Se μ fosse conhecida, um estimador não-viesado de σ^2 seria $n^{-1}\sum_{i=1}^{n}(Y_i - \mu)^2$, mas na prática μ é raramente conhecida.

Embora a inexistência de viés tenha um certo apelo como uma propriedade de um estimador — de fato, seu antônimo, "viesado", tem decididamente conotações negativas —, ela não está livre de problemas. Um ponto fraco da inexistência de viés é que alguns estimadores razoáveis, e até mesmo muito bons, são viesados. Brevemente veremos um exemplo.

Um outro ponto fraco importante da inexistência de viés é que existem estimadores não-viesados que de fato são estimadores bastante pobres. Considere estimar a média μ de uma população. Em lugar de usar a média amostral $\bar{Y}$ para estimar μ, suponha que, após coletar uma amostra de tamanho n, descartemos todas as observações, exceto a primeira. Ou seja, nosso estimador de μ será simplesmente $W \equiv Y_1$. Esse estimador será não-viesado, pois $E(Y) = \mu$. Esperançosamente, você perceberá que ignorar todas as observações, exceto a primeira, não é um método prudente de estimação: ele joga fora a maioria das informações da amostra. Por exemplo, com $n = 100$, obteremos 100 resultados da variável aleatória Y, mas usaremos somente a primeira delas para estimar $E(Y)$.

A Variância Amostral dos Estimadores

O exemplo no final da subseção anterior mostra que precisamos de critérios adicionais para avaliar os estimadores. A inexistência de viés somente garante que a distribuição amostral de um estimador tem um valor médio igual ao parâmetro que ela supostamente está estimando. Isso é bom, mas também precisamos saber o quanto está espalhada a distribuição de um estimador. Um estimador pode ser igual a θ, em média, mas também pode estar muito longe com probabilidade grande. Na Figura C.2, W_1 e W_2 são ambos estimadores não-viesados de θ. Contudo, a distribuição de W_1 está mais firmemente centralizada em torno de θ: a probabilidade de W_1 ser maior que qualquer determinada distância de θ é menor que a probabilidade de W_2 ser maior que a mesma distância de θ. O uso da W_1 como nosso estimador significa que é menos provável que venhamos a obter uma amostra aleatória que produza uma estimativa muito afastada de θ.

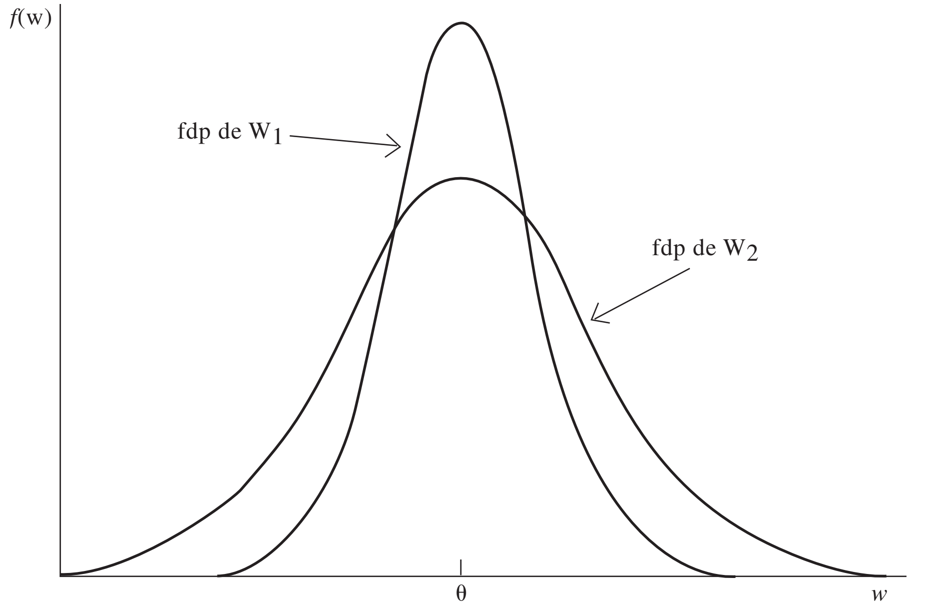
Para resumir a situação mostrada na Figura C.2, apoiamo-nos na variância (ou desvio-padrão) de um estimador. Recorde-se que isso produz uma medida única da dispersão na distribuição. A variância de um estimador é freqüentemente chamada de **variância amostral**, pois ela é a variância associada a uma distribuição amostral. Lembre-se, a variância amostral não é uma variável aleatória; ela é uma constante, mas pode ser desconhecida.

Agora obteremos a variância da média amostral para estimar a média μ de uma população:

$$\begin{aligned} \mathrm{Var}(\bar{Y}) &= \mathrm{Var}\left((1/n)\sum_{i=1}^{n}Y_i\right) = (1/n^2)\mathrm{Var}\left(\sum_{i=1}^{n}Y_i\right) = (1/n^2)\left(\sum_{i=1}^{n}\mathrm{Var}(Y_i)\right) \\ &= (1/n^2)\left(\sum_{i=1}^{n}\sigma^2\right) = (1/n^2)(n\sigma^2) = \sigma^2/n. \end{aligned} \tag{C.6}$$

Figura C.2

Distribuições amostrais de dois estimadores não-viesados de θ.



Observe como usamos as propriedades da variância das Seções B.3 e B.4 (VAR.2 e VAR.4), assim como a independência dos Y_i. Para resumir: se $\{Y_i$: $i = 1, 2, ..., n)$ for uma amostra aleatória de uma população com média μ e variância σ^2, então, $\bar{Y}$ terá a mesma média da população, mas sua variância amostral será igual à variância populacional, σ^2, dividida pelo tamanho da amostra.

Uma implicação importante de $\text{Var}(\bar{Y}) = \sigma^2/n$ é que ela pode ficar muito próxima de zero aumentando do tamanho da amostra n. Essa é uma característica-chave de um estimador razoável, e voltaremos a ele na Seção C.3.

Como sugerido pela Figura C.2, entre os estimadores não-viesados, preferimos o estimador com a menor variância. Isso nos possibilita desconsiderar certos estimadores. Para uma amostra aleatória com média μ e variância σ^2, sabemos que $\bar{Y}$ será não-viesado, e $\text{Var}(\bar{Y}) = \sigma^2/n$. E quanto ao estimador Y_1, que é simplesmente a primeira observação extraída? Como Y_1 é uma extração aleatória da população, $\text{Var}(Y_1) = \sigma^2$. Assim, a diferença entre $\text{Var}(Y_1)$ e $\text{Var}(\bar{Y})$ poderá ser grande mesmo para amostras de tamanhos pequenos. Se $n = 10$, então, $\text{Var}(Y_1)$ será dez vezes maior que $\text{Var}(\bar{Y}) = \sigma^2/10$. Isso nos oferece uma maneira formal para excluir Y_1 como um estimador de μ.

Para enfatizar esse ponto, a Tabela C.1 contém o resultado de um pequeno estudo simulado. Usando o programa estatístico Stata, 20 amostras aleatórias de tamanho 10 foram geradas a partir de uma distribuição normal, com $\mu = 2$ e $\sigma^2 = 1$; nesse caso, estamos interessados em estimar μ. Para cada uma das 20 amostras aleatórias, computamos duas estimativas, y_1 e $\bar{y}$; esses valores estão descritos na Tabela C.1. Como pode ser visto na tabela, os valores de y_1 são muito mais dispersos que os de $\bar{y}$: y_1 varia de $-0{,}64$ a 4,27, enquanto $\bar{y}$ varia somente de 1,16 a 2,58. Além disso, em 16 dos 20 casos, $\bar{y}$ está mais próximo de $\mu = 2$ que y_1. A média de y_1 na simulação está em torno de 1,89, enquanto a de $\bar{y}$ é de 1,96. O fato de que essas médias estão próximas de 2 ilustra a inexistência de viés de ambos os estimadores (e poderíamos obter essas médias mais próximas de 2 se utilizássemos mais de 20 amostras). Mas a comparação apenas dos resultados médios entre as extrações aleatórias mascara o fato de que a média amostral $\overline{Y}$ é muito superior a Y_1 como um estimador de μ.

Tabela C.1

Simulação de Estimadores para uma Distribuição Normal(μ,1) com $\mu = 2$

Amostra	y_1	$\bar{y}$
1	−0,64	1,98
2	1,06	1,43
3	4,27	1,65
4	1,03	1,88
5	3,16	2,34
6	2,77	2,58
7	1,68	1,58
8	2,98	2,23
9	2,25	1,96
10	2,04	2,11
11	0,95	2,15
12	1,36	1,93
13	2,62	2,02
14	2,97	2,10
15	1,93	2,18
16	1,14	2,10
17	2,08	1,94
18	1,52	2,21
19	1,33	1,16
20	1,21	1,75

Eficiência

A comparação das variâncias de $\overline{Y}$ e Y_1 na subseção anterior é um exemplo de um método genérico para comparar diferentes estimadores não-viesados.

EFICIÊNCIA RELATIVA

Se W_1 e W_2 forem dois estimadores não-viesados de θ, W_1 será eficiente com relação a W_2 quando $\text{Var}(W_1) \leq \text{Var}(W_2)$ para qualquer θ, com desigualdade estrita para pelo menos um valor de θ.

Anteriormente, mostramos que, para estimar a média populacional μ, $\text{Var}(\bar{Y}) < \text{Var}(Y_1)$ para qualquer valor de σ^2 sempre que $n > 1$. Assim, $\bar{Y}$ é eficiente em relação a Y_1 para estimar μ. Não podemos sempre escolher entre os estimadores não-viesados com base no critério de menor variância: dados dois estimadores não-viesados de μ, um poderá ter menor variância para alguns valores de μ, enquanto o outro poderá ter menor variância para outros valores de μ.

Se restringirmos nossa atenção para certa classe de estimadores, poderemos mostrar que a média amostral tem a menor variância. O Problema C.2 pede que você mostre que $\bar{Y}$ tem a menor variância entre todos os estimadores não-viesados que também sejam funções lineares de $Y_1, Y_2, ..., Y_n$. As hipóteses são que os Y_i têm média e variância comuns, e que eles são não-correlacionadas dois a dois.

Se não restringirmos nossa atenção aos estimadores não-viesados, então, não terá sentido comparar as variâncias. Por exemplo, quando estimamos a média populacional μ, podemos usar um estimador trivial que seja igual a zero, independente da amostra extraída. Naturalmente, a variância desse estimador será zero (já que será o mesmo valor para qualquer amostra aleatória). Porém, o viés desse estimador será $-\mu$, e assim ele será um estimador muito pobre quando $|\mu|$ for grande.

Uma maneira de comparar estimadores que não sejam necessariamente não-viesados é calcular o **erro quadrático médio** (**EQM**) dos estimadores. Se W for um estimador de θ, então, o EQM de W será definido como $\text{EQM}(W) = \text{E}[(W - \mu)^2]$. O EQM mede o quanto o estimador está distante, em média, de μ. É possível mostrar que $\text{EQM}(W) = \text{Var}(W) + [\text{Viés}(W)]^2$, de forma que $\text{EQM}(W)$ depende da variância e do viés (se algum estiver presente). Isso nos possibilita comparar dois estimadores quando houver viés em um ou ambos.

C.3 PROPRIEDADES ASSIMPTÓTICAS DOS ESTIMADORES OU PROPRIEDADES DOS ESTIMADORES EM AMOSTRAS GRANDES

Na Seção C.2, encontramos o estimador Y_1 da média populacional μ, e vimos que, embora ele fosse não-viesado, era um estimador pobre, pois sua variância poderia ser muito maior que a da média amostral. Uma característica notável de Y_1 é que ele tem a mesma variância para qualquer tamanho de amostra. Parece razoável exigir que qualquer procedimento de estimação se aprimore conforme o tamanho da amostra se torne maior. Para estimar uma média populacional μ, $\bar{Y}$ melhora no sentido de que sua variância vai se tornando menor conforme n vai ficando maior; Y_1 não melhora nesse sentido.

Podemos excluir certos estimadores absurdos estudando as propriedades *assimptóticas* ou de *amostras grandes* dos estimadores. Além disso, podemos dizer alguma coisa positiva sobre os estimadores que não são não-viesados e cujas variâncias não são encontradas com facilidade.

A análise assimptótica envolve a aproximação das características da distribuição amostral de um estimador. Essas aproximações dependem do tamanho da amostra. Infelizmente, estamos obrigatoriamente limitados quanto ao que podemos dizer com respeito a o quanto uma amostra precisa ser "grande" para que uma análise assimptótica seja apropriada; isso depende da distribuição populacional subjacente. Porém, aproximações de amostras grandes têm se mostrado funcionar bem para tamanhos de amostras tão pequenas quanto $n = 20$.

Consistência

A primeira propriedade assimptótica dos estimadores se refere à provável distância que o estimador fica do parâmetro que ele supostamente estará estimando conforme o tamanho da amostra cresça indefinidamente.

CONSISTÊNCIA

Seja W_n um estimador de θ com base em uma amostra $Y_1, Y_2, ..., Y_n$ de tamanho n. Então, W_n será um **estimador consistente** de θ se, para cada $\varepsilon > 0$,

$$P(|W_n - \theta| > \varepsilon) \to 0 \text{ conforme } n \to \infty. \quad \textbf{(C.7)}$$

Se W_n não for consistente para θ, então, dizemos que ele é **inconsistente**.

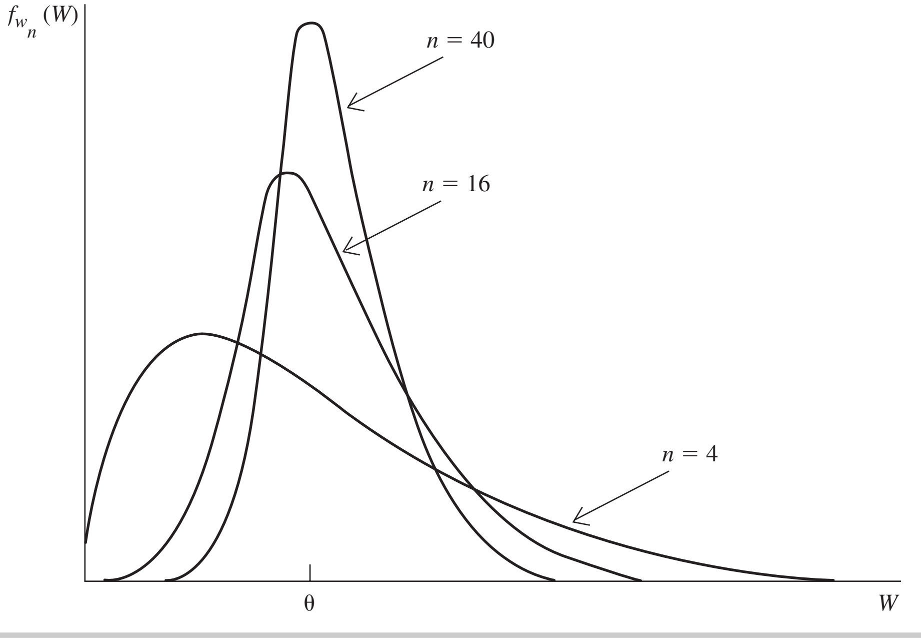
Quando W_n é consistente, também dizemos que θ é o **limite de probabilidade** de W_n, escrito como $\text{plim}(W_n) = \theta$.

Ao contrário da inexistência de viés — que é uma característica de um estimador para um determinado tamanho de amostra —, a consistência envolve o comportamento da distribuição amostral do estimador conforme o tamanho da amostra n fica maior. Para destacar isso, indexamos o estimador pelo tamanho da amostra declarando essa definição, e continuaremos com essa convenção por toda esta seção.

A equação (C.7) parece técnica e pode ser muito difícil de ser determinada com base nos princípios fundamentais da probabilidade. Por outro lado, a interpretação da (C.7) é direta. Ela significa que a distribuição de W_n se torna cada vez mais concentrada em torno de θ, o que grosso modo significa que, para amostras de tamanhos maiores, será cada vez menos provável que W_n fique muito afastado de θ. Essa tendência está ilustrada na Figura C.3.

Figura C.3

As distribuições amostrais de um estimador consistente para três tamanhos de amostra.



Se um estimador não for consistente, ele não nos ajudará na obtenção de informações sobre θ, mesmo com uma quantidade ilimitada de dados. Por essa razão, a consistência é um requisito mínimo de um estimador usado em estatística ou econometria. Encontraremos estimadores que são consistentes sob certas hipóteses e inconsistentes quando essas hipóteses falham. Quando os estimadores são inconsistentes, em geral podemos encontrar seus limites de probabilidade, e será importante saber o quanto esses limites de probabilidade estão distantes de θ.

Como observamos antes, estimadores não-viesados não são necessariamente consistentes, mas aqueles cujas variâncias tendem para zero conforme o tamanho da amostra cresce *são* consistentes. Isso pode ser estabelecido formalmente: se W_n for um estimador não-viesado de θ e $\text{Var}(W_n) \to \infty$ conforme $n \to \infty$, então, $\text{plim}(W_n) = \theta$. Estimadores não-viesados que usam a totalidade da amostra de dados geralmente terão uma variância que se reduzirá para zero conforme o tamanho da amostra cresça, sendo, portanto, consistentes.

Um bom exemplo de um estimador consistente é a média de uma amostra aleatória extraída de uma população com média μ e variância σ^2. Já mostramos que a média amostral é não-viesada para μ. Na equação (C.6), derivamos $\text{Var}(\bar{Y}_n) = \sigma^2/n$ para qualquer amostra de tamanho n. Portanto, $\text{Var}(\bar{Y}_n) \to 0$ conforme $n \to \infty$, e, portanto, $\bar{Y}_n$ é um estimador consistente de μ (além de ser não-viesado).

A conclusão que $\bar{Y}_n$ é consistente para μ é válida mesmo se $\text{Var}(\bar{Y}_n)$ não existir. Esse resultado clássico é conhecido como a **lei dos grandes números (LGN)**.

LEI DOS GRANDES NÚMEROS

Sejam $Y_1, Y_2, ..., Y_n$ variáveis aleatórias independentes e identicamente distribuídas com média μ. Então,

$$\text{plim}(\bar{Y}_n) = \mu. \tag{C.8}$$

A lei dos grandes números significa que, se estivermos interessados em estimar a média populacional μ, poderemos chegar arbitrariamente próximos de μ, escolhendo uma amostra suficientemente grande. Esse resultado fundamental pode ser combinado com propriedades básicas dos limites de probabilidade para mostrar que estimadores razoavelmente complicados são consistentes.

PROPRIEDADE PLIM.1

Seja θ um parâmetro e defina um novo parâmetro, $\gamma = g(\theta)$ para alguma função *contínua* $g(\theta)$. Suponha que $\text{plim}(W_n) = \theta$. Defina um estimador de γ como $G_n = g(W_n)$. Então,

$$\text{plim}(G_n) = \gamma. \tag{C.9}$$

Isso é em geral definido como

$$\text{plim } g(W_n) = g(\text{plim } W_n) \tag{C.10}$$

para uma função contínua $g(\theta)$.

A hipótese de que $g(\theta)$ é contínua é um requisito técnico que freqüentemente tem sido descrito de forma não técnica como "uma função que pode ser traçada sem precisar levantar o lápis do papel". Como todas as funções que encontramos neste livro são contínuas, não apresentamos uma definição formal de função contínua. São exemplos de funções contínuas $g(\theta) = a + b\theta$ para constantes a e b, $g(\theta) = \theta^2$, $g(\theta) = 1/\theta$, $g(\theta) = \sqrt{\theta}$, $g(\theta) = \exp(\theta)$, e muitas outras variantes destas. Não precisaremos citar novamente a hipótese de continuidade.

Como um exemplo importante de um estimador consistente, mas viesado, considere estimar o desvio-padrão, σ, de uma população com média μ e variância σ^2. Já afirmamos que a variância amostral $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ é não-viesada para σ^2. Usando a lei dos grandes números e um pouco de álgebra, também é possível mostrar que S_n^2 é consistente para σ^2. O estimador natural de $\sigma = \sqrt{\sigma^2}$ é $S_n = \sqrt{S_n^2}$ (onde a raiz quadrada é sempre a raiz quadrada positiva). S_n, que é chamado **desvio-padrão amostral**, *não* é um estimador não-viesado porque o valor esperado da raiz quadrada *não* é a raiz quadrada do valor esperado (veja a Seção B.3). No entanto, de acordo com a PLIM.1, plim $S_n = \sqrt{\text{plim } S_n^2}$ $= \sqrt{\sigma^2} = \sigma$, de modo que S_n é um estimador consistente de σ.

A seguir algumas outras propriedades úteis do limite de probabilidade:

PROPRIEDADE PLIM.2

Se $\text{plim}(T_n) = \alpha$ e $\text{plim}(U_n) = \beta$, então,

(i) $\text{plim}(T_n + U_n) = \alpha + \beta$;

(ii) $\text{plim}(T_nU_n) = \alpha\beta$;

(iii) $\text{plim}(T_n/U_n)\ \alpha/\beta$, desde que $\beta \neq 0$.

Esses três fatos sobre os limites de probabilidade nos possibilitam combinar estimadores consistentes de várias maneiras para obter outros estimadores consistentes. Por exemplo, seja $\{Y_1, ..., Y_n\}$ uma amostra aleatória de tamanho n dos rendimentos anuais da população de trabalhadores com ensino médio completo e seja a média populacional dada por μ_Y. Seja $\{Z_1, ..., Z_n\}$ uma amostra aleatória dos rendimentos anuais da população de trabalhadores com curso superior completo e seja a média populacional dada por μ_Z. Queremos estimar a diferença porcentual dos rendimentos anuais entre os dois grupos, que será $\gamma = 100\cdot(\mu_Z - \mu_Y)/\mu_Y$. (Essa será a porcentagem pela qual os rendimentos médios daqueles com curso superior diferirão dos rendimentos médios daqueles com ensino médio). Como $\bar{Y}_n$ é consistente para μ_Y, e $\bar{Z}_n$ é consistente para μ_Z, decorre de PLIM.1 e da parte (iii) de PLIM.2 que

$$G_n \equiv 100\cdot(\bar{Z}_n - \bar{Y}_n)/\bar{Y}_n$$

é um estimador consistente de γ. G_n é simplesmente a diferença percentual entre $\bar{Z}_n$ e $\bar{Y}_n$ na amostra, de modo que ele é um estimador natural. G_n não é um estimador não-viesado de γ, mas ainda assim é um bom estimador, a menos que n seja pequeno.

Normalidade Assimptótica

Consistência é uma propriedade dos estimadores por ponto. Embora ela nos informe que a distribuição do estimador está se concentrando em torno do parâmetro conforme o tamanho da amostra vai ficando maior, ela nada nos diz sobre a *forma* daquela distribuição para uma amostra de determinado tamanho. Para construir estimadores por intervalo e para testar hipóteses, precisamos de uma maneira de aproximar a distribuição de nossos estimadores. A maioria dos estimadores econométricos possui distribuições que são bem aproximadas por uma distribuição normal para amostras grandes, o que motiva a seguinte definição.

NORMALIDADE ASSIMPTÓTICA

Seja $\{Z_n: n = 1, 2, ...\}$ uma seqüência de variáveis aleatórias, de forma que para todos os números z,

$$P(Z_n \leq z) \to \Phi(z) \text{ conforme } n \to \infty, \tag{C.11}$$

onde $\Phi(z)$ é a função de distribuição cumulativa normal padrão. Então, diz-se que Z_n tem um *distribuição normal padrão assimptótica*. Nesse caso, freqüentemente escrevemos $Z_n \overset{a}{\sim}$ Normal(0,1). (O "a" acima do til significa "assimptoticamente" ou "aproximadamente").

A Propriedade (C.11) significa que a função de distribuição cumulativa de Z_n se aproxima cada vez mais da fdc da distribuição normal padrão conforme o tamanho n da amostra vai ficando maior. Quando a **normalidade assimptótica** é válida, teremos, para n grande, a aproximação $P(Z_n \leq z) \approx \Phi(z)$. Dessa forma, as probabilidades concernentes a Z_n poderão ser aproximadas pelas probabilidades normais padrões.

O **teorema do limite central** (**TLC**) é um dos resultados mais poderosos em probabilidade e estatística. Ele afirma que a média de uma amostra aleatória de *qualquer* população (com variância finita), quando padronizada, tem uma distribuição normal padrão assimptótica.

TEOREMA DO LIMITE CENTRAL

Seja $\{Y_1, Y_2, ..., Y_n\}$ uma amostra aleatória com média μ e variância σ^2. Então,

$$Z_n = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \tag{C.12}$$

tem uma distribuição normal padrão assimptótica.

A variável Z_n em (C.12) é a versão padronizada de $\bar{Y}_n$: subtraímos $E(\bar{Y}_n) = \mu$ e dividimos por $dp(\bar{Y}_n) = \sigma/\sqrt{n}$. Dessa forma, independentemente da distribuição populacional de Y, Z_n terá média zero e variância um, que coincide com a média e a variância da distribuição normal padrão. Notadamente, a totalidade da distribuição de Z_n se aproxima arbitrariamente da distribuição normal padrão conforme n vai ficando maior.

Podemos escrever a variável padronizada na equação (C.12) como $\sqrt{n}(\bar{Y}_n - \mu)/\sigma$, que mostra que devemos multiplicar a diferença entre a média amostral e a média populacional pela raiz quadrada do tamanho da amostra, para obtermos uma distribuição limitada proveitosa. Sem a multiplicação por $\sqrt{n}$, apenas teríamos $(\bar{Y}_n - \mu)/\sigma$, que converge em probabilidade para zero. Em outras palavras, a distribuição de $(\bar{Y}_n - \mu)/\sigma$ simplesmente cai para um único ponto conforme $n \to \infty$, que sabemos não poder ser uma boa aproximação para a distribuição de $(\bar{Y}_n - \mu)/\sigma$ para amostras de tamanhos razoáveis. A multiplicação por $\sqrt{n}$ garante que a variância de Z_n permaneça constante. Na prática, é comum tratarmos $\bar{Y}_n$ como aproximadamente normalmente distribuída com média μ e variância σ^2/n, e isso nos dá os procedimentos estatísticos corretos, pois leva à variável padronizada na equação (C.12).

A maioria dos estimadores encontrados em estatística e econometria pode ser escrita como funções de médias amostrais, caso em que podemos aplicar a lei dos grandes números e o teorema do limite central. Quando dois estimadores consistentes têm distribuições normais assimptóticas, selecionamos o estimador com a menor variância assimptótica.

Além da média amostral padronizada em (C.12), muitas outras estatísticas que dependem de médias amostrais acabam sendo assimptoticamente normais. Uma estatítica importante é obtida pela substituição de σ por seu estimador consistente S_n na equação (C.12):

$$\frac{\bar{Y}_n - \mu}{S_n/\sqrt{n}} \tag{C.13}$$

também tem uma distribuição normal padrão aproximada para n grande. As distribuições exatas (amostra finita) de (C.12) e (C.13) não são, definitivamente, as mesmas, mas a diferença será com freqüência pequena o suficiente para ser ignorada para n grande.

Em toda esta seção, cada estimador tem tido um subscrito n para enfatizar a natureza da análise assimptótica ou de amostra grande. A continuação dessa convenção confundirá a notação sem fornecer informações adicionais, uma vez que os fundamentos da análise assimptótica tenham sido compreendidos. De agora em diante, eliminaremos o subscrito n e confiaremos que você se lembrará que os estimadores dependem do tamanho da amostra, e que propriedades como consistência e normalidade assimptótica referem-se ao crescimento do tamanho da amostra sem limites.

C.4 MÉTODOS GERAIS DE ESTIMAÇÃO DE PARÂMETROS

Até aqui, usamos a média amostral para ilustrar as propriedades dos estimadores finitos e de amostras grandes. É natural perguntar se há métodos gerais de estimação que produzem estimadores com boas propriedades, tais como a inexistência de viés, consistência e eficiência.

A resposta é sim. Uma abordagem detalhada de vários métodos de estimação está além do escopo deste trabalho; aqui apresentamos somente uma discussão informal. Um exame completo é feito por Larsen e Marx (1986, Capítulo 5).

Método dos Momentos

Dado um parâmetro θ aparecendo em uma distribuição populacional, usualmente existem muitas maneiras para obter estimadores não-viesados e consistentes de θ. Tentar todas as diferentes possibilidades e compará-las com base nos critérios das seções C.2 e C.3 não é prático. Felizmente, alguns métodos têm mostrado ter boas propriedades gerais e, na maior parte, a lógica deles é intuitivamente interessante.

Nas seções anteriores, estudamos a média amostral como um estimador não-viesado da média populacional e a variância amostral como um estimador não-viesado da variância populacional. Esses estimadores são exemplos de estimadores obtidos pelo **método dos momentos**. De forma geral, a estimação pelo método dos momentos é feita da seguinte maneira: o parâmetro θ é mostrado como relacionado com algum valor esperado na distribuição de Y, em geral $E(Y)$ ou $E(Y^2)$ (embora algumas vezes sejam usadas escolhas menos comuns). Suponha, por exemplo, que o parâmetro de interesse, θ, seja relacionado com a média populacional como $\theta = g(\mu)$ para alguma função g. Como a média amostral $\bar{Y}$ é um estimador não-viesado e consistente de μ, é natural substituir μ por $\bar{Y}$, o que nos dará o estimador $g(\bar{Y})$ de θ. O estimador $g(\bar{Y})$ será consistente para θ, e se $g(\mu)$ for uma função linear de μ, então, $g(\bar{Y})$ também será não-viesado. O que fizemos foi substituir o momento populacional, μ, por seu equivalente amostral, $\bar{Y}$. É daí que vem o nome "método dos momentos".

Tratamos de mais dois estimadores pelo método dos momentos que serão úteis para nossa discussão sobre a análise de regressão. Recorde-se de que a covariância entre duas variáveis aleatórias

X e Y é definida como $\sigma_{XY} = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$. O método dos momentos sugere estimar σ_{XY} por $n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$. Este será um estimador consistente de σ_{XY}, mas ele será viesado essencialmente pela mesma razão que a variância amostral será viesada se n, em lugar de $n - 1$, for usado como divisor. A **covariância amostral** é definida como

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}). \tag{C.14}$$

É possível mostrar que este é um estimador não-viesado de σ_{XY}. (A substituição de n por $n - 1$ não faz diferença se o tamanho da amostra crescer indefinidamente, de modo que este estimador ainda será consistente.)

Como discutimos na seção B.4, a covariância entre duas variáveis muitas vezes é difícil de ser interpretada. Em geral, estamos mais interessados na correlação. Como a correlação populacional é $\rho_{XY} = \sigma_{XY}/(\sigma_X\sigma_Y)$, o método dos momentos sugere estimar ρ_{XY} como

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{1/2}\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)^{1/2}}, \tag{C.15}$$

que é chamado **coeficiente de correlação amostral** (ou, abreviadamente, correlação amostral). Observe que cancelamos a divisão por $n - 1$ na covariância amostral e nos desvios-padrão amostrais. Na realidade, poderíamos dividir cada um deles por n e chegar na mesma fórmula final.

É possível mostrar que o coeficiente de correlação amostral estará sempre no intervalo $[-1,1]$, como deveria ser. Como S_{XY}, S_X e S_Y são consistentes em relação ao parâmetro populacional correspondente, R_{XY} é um estimador consistente da correlação populacional, ρ_{XY}. Entretanto, R_{XY} será um estimador viesado por duas razões. Primeiro, S_X e S_Y são estimadores viesados de σ_X e σ_Y, respectivamente. Segundo, R_{XY} é uma razão de estimadores, e assim ele não seria não-viesado, mesmo se S_X e S_Y fossem. Para nosso propósito, isso não é importante, embora o fato de não existir um estimador não-viesado de ρ_{XY} seja um resultado clássico em estatística matemática.

Máxima Verossimilhança

Outro método geral de estimação é o da *máxima verossimilhança*, um assunto tratado em muitos cursos introdutórios de estatística. Um breve resumo do caso mais simples será suficiente aqui. Seja $\{Y_1, Y_2, ..., Y_n\}$ uma amostra aleatória da distribuição populacional $f(y;\theta)$. Devido à hipótese de amostragem aleatória, a distribuição conjunta de $\{Y_1, Y_2, ..., Y_n\}$ será simplesmente o produto das densidades: $f(y_1;\theta)f(y_2;\theta) \cdots f(y_n;\theta)$. No caso discreto, isso será $\mathrm{P}(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n)$. Agora, defina a *função de verossimilhança* como

$$V(\theta;Y_1, ...,Y_n) = f(Y_1;\theta)f(Y_2;\theta) \cdots f(Y_n;\theta),$$

que será uma variável aleatória, pois ela depende do resultado da amostra aleatória $\{Y_1, Y_2, ..., Y_n\}$. O **estimador de máxima verossimilhança** de θ, vamos chamá-lo de W, será o valor de θ que maximiza a função de verossimilhança. (Esse é o motivo pelo qual escrevemos V como uma função de θ, seguido da amostra aleatória). Claramente, esse valor depende da amostra aleatória. O princípio da máxima verossimilhança diz que, de todos os valores possíveis de θ, o valor que torna a verossimilhança do dado observado a maior deve ser escolhido. Intuitivamente, esse é um método razoável de estimar θ.

Geralmente, é mais conveniente trabalhar com a *função log-verossimilhança*, que é obtida tomando-se o log natural da função de verossimilhança:

$$\log[V(\theta; Y_{1,\ \ldots,}\ Y_n)] = \sum_{i=1}^{n} \log\,[f(Y_i;\,\theta)], \tag{C.16}$$

quando usamos o fato de que o log do produto é a soma dos logs. Como (C.16) é a soma de variáveis aleatórias independentes e identicamente distribuídas, analisar os estimadores que decorrem de (C.16) é relativamente fácil.

A estimação da máxima verossimilhança (EMV) em geral é consistente e algumas vezes não-viesada. Mas também o são muitos outros estimadores. A atração da EMV é que ela geralmente fornece o estimador mais assimptoticamente eficiente quando o modelo populacional $f(y;\theta)$ é corretamente especificado. Além disso, a EMV algumas vezes é o **estimador não-viesado de variância mínima**; isto é, ela tem a menor variância entre os estimadores não-viesados de θ. [Veja Larsen e Marx (1986, Capítulo 5) para verificar essas afirmações.]

No Capítulo 17, precisaremos da máxima verossimilhança para estimar os parâmetros de modelos econométricos mais avançados. Em econometria, estamos quase sempre interessados na distribuição de Y condicional a um conjunto de variáveis explicativas, digamos X_1, X_2, ..., X_k. Depois, substituímos a densidade em (C.16) por $f(Y_1|X_{i1}, \ldots, X_{ik}; \theta_1, \ldots, \theta_p)$, onde é permitida a essa densidade depender de p parâmetros, θ_1, ..., θ_p. Felizmente, para aplicações bem-sucedidas de métodos de máxima verossimilhança, não precisamos nos aprofundar muito nos problemas computacionais ou na teoria estatística de amostras grandes. Wooldridge (2002, Capítulo 13) trata da teoria da estimação por máxima verossimilhança.

Mínimos Quadrados

Um terceiro tipo de estimador, e um que tem um papel importante ao longo de todo este livro, é chamado de **estimador de mínimos quadrados**. Já vimos um exemplo de mínimos quadrados: a média amostral $\bar{Y}$ é um estimador de mínimos quadrados da média populacional, μ. Já sabemos que $\bar{Y}$ é um estimador pelo método dos momentos. O que o torna um estimador de mínimos quadrados? É possível mostrar que o valor de m que torna a soma dos desvios quadrados

$$\sum_{i=1}^{n} (Y_i - m)^2$$

tão pequena quanto possível é $m = \bar{Y}$. Demonstrar isso não é difícil, mas omitiremos a álgebra.

Para algumas distribuições importantes, inclusive a normal e a de Bernoulli, a média amostral $\bar{Y}$ também é o estimador de máxima verossimilhança da média populacional μ. Assim, os princípios dos mínimos quadrados, do método dos momentos e da máxima verossimilhança muitas vezes resultam no *mesmo* estimador. Em outros casos, os estimadores são semelhantes, mas não idênticos.

C.5 ESTIMAÇÃO POR INTERVALO E INTERVALOS DE CONFIANÇA

A Natureza da Estimação por Intervalo

Uma estimativa por ponto obtida a partir de uma amostra particular não fornece, por si só, informações suficientes para testar teorias econômicas ou para explicar detalhes de decisões. Uma estimativa por ponto poderá ser a melhor suposição do pesquisador do valor da população, mas, por sua natureza, ela não fornece informação sobre o quanto é "provável" que ela deva estar próxima do parâmetro populacional. Como um exemplo, suponha que um pesquisador descreva, com base em uma amostra aleatória de trabalhadores, que os subsídios de treinamento de pessoal aumentam o salário por hora em 6,4%. Como poderemos saber se ou não esse número está próximo do efeito na população de trabalhadores que podem ter sido treinados? Como não conhecemos o valor da população, não podemos saber o quanto está próxima uma estimativa de determinada amostra. Porém, podemos fazer afirmações envolvendo probabilidades, e é aqui que entra a estimação por intervalo.

Já conhecemos uma maneira de avaliar a incerteza em um estimador: encontre seu desvio-padrão amostral. Informar o desvio-padrão do estimador, com a estimativa por ponto, fornece alguma informação sobre a precisão de nossa estimativa. Porém, mesmo se o problema da dependência do desvio-padrão em relação a parâmetros populacionais desconhecidos for ignorada, informar o desvio-padrão do estimador, com a estimativa por ponto, não significa qualquer afirmação direta sobre onde o valor da população provavelmente estará em relação à estimativa. Essa limitação é superada pela construção de um **intervalo de confiança**.

Ilustramos o conceito de um intervalo de confiança com um exemplo. Suponha que a população tem uma distribuição Normal(μ,1) e seja $\{Y_1, ..., Y_n\}$ uma amostra aleatória dessa população. (Assumimos que a variância da população é conhecida e igual a unidade para o fim desta ilustração; depois mostraremos o que fazer no caso mais real em que a variância é desconhecida.) A média amostral, $\bar{Y}$, tem uma distribuição normal com média μ e variância $1/n$: $\bar{Y} \sim \text{Normal}(\mu,1/n)$. A partir daí podemos padronizar $\bar{Y}$, e como a versão padronizada de $\bar{Y}$ tem uma distribuição normal padrão, teremos

$$\text{P}\left(-1{,}96 < \frac{\bar{Y} - \mu}{1/\sqrt{n}} < 1{,}96\right) = 0{,}95.$$

O evento entre parênteses é idêntico ao evento $\bar{Y} - 1{,}96/\sqrt{n} < \mu < \bar{Y} + 1{,}96/\sqrt{n}$, e, portanto,

$$\text{P}(\bar{Y} - 1{,}96/\sqrt{n} < \mu < \bar{Y} + 1{,}96/\sqrt{n}) = 0{,}95. \tag{C.17}$$

A equação (C.17) é interessante por nos informar que a probabilidade de o intervalo aleatório $[\bar{Y} - 1{,}96/\sqrt{n}, \bar{Y} + 1{,}96/\sqrt{n}]$ conter a média populacional μ é de 0,95, ou 95%. Essa informação nos permite construir uma *estimativa por intervalo* de μ, que é obtida pela agregação do resultado amostral da média, $\bar{y}$. Assim,

$$[\bar{y} - 1{,}96/\sqrt{n}, \bar{y} + 1{,}96/\sqrt{n}] \tag{C.18}$$

é um exemplo de uma estimativa por intervalo de μ. Ela também é chamada de intervalo de confiança de 95%. Uma notação abreviada desse intervalo é $\bar{y} \pm 1{,}96/\sqrt{n}$.

É fácil calcular o intervalo de confiança na equação (C.18), logo que os dados da amostra $\{y_1, y_2, ..., y_n\}$ sejam observados; $\bar{y}$ é o único fator que depende dos dados. Por exemplo, suponha que $n = 16$ e que a média dos 16 pontos de dados seja 7,3. Então, o intervalo de confiança de 95% de μ será 7,3 $\pm$ 1,96/$\sqrt{16}$ = 7,3 $\pm$ 0,49, que podemos escrever na forma de intervalo como [6,81; 7,79]. Por construção, $\bar{y} = 7{,}3$ está no centro desse intervalo.

Ao contrário de seu cálculo, o significado de um intervalo de confiança é mais difícil de entender. Quando dizemos que a equação (C.18) é um intervalo de confiança de 95% de μ, queremos dizer que o intervalo *aleatório*

$$[\bar{Y} - 1{,}96/\sqrt{n},\ \bar{Y} + 1{,}96/\sqrt{n}] \tag{C.19}$$

contém μ com uma probabilidade de 0,95. Em outras palavras, *antes* de extrair a amostra aleatória, existe 95% de possibilidade de que (C.19) contenha μ. A equação (C.19) é um exemplo de um **estimador por intervalo**. Ele é um intervalo aleatório, pois as extremidades mudam com diferentes amostras.

Um intervalo de confiança muitas vezes é interpretado da seguinte maneira: "a probabilidade de que μ esteja no intervalo (C.18) é de 95%". Isso é incorreto. Uma vez que a amostra tenha sido observada e $\bar{y}$ tenha sido calculado, os limites do intervalo de confiança serão simplesmente números (6,81 e 7,79 no exemplo dado). O parâmetro populacional, μ, embora desconhecido, também será apenas algum número. Portanto, μ estará ou não no intervalo (C.18) (e nunca saberemos com certeza se isso é verdadeiro). A probabilidade não desempenha papel algum, uma vez que o intervalo de confiança tenha sido calculado para os dados disponíveis. A interpretação probabilística advém do fato de que, para 95% de todas as amostras aleatórias, o intervalo de confiança construído contém μ.

Para destacar o significado de um intervalo de confiança, a Tabela C.2 contém cálculos para 20 amostras aleatórias da distribuição Normal(2,1) com amostras de tamanho $n = 10$. Para cada uma das 20 amostras, $\bar{y}$ é obtido, e (C.18) é calculado como $\bar{y} \pm 1{,}96/\sqrt{10} = \bar{y} \pm 0{,}62$ (cada qual arredondado para duas casas decimais). Como é possível ver, o intervalo muda com cada amostra aleatória. Dezenove dos 20 intervalos contêm o valor populacional de μ. Somente na amostra número 19 μ não está no intervalo de confiança. Em outras palavras, 95% das amostras resultam em um intervalo de confiança que contém μ. Nem sempre esse é o caso com somente 20 amostras, mas funcionou dessa maneira nessa simulação em particular.

Tabela C.2

Intervalos de Confiança Simulados para uma Distribuição Normal(μ,1) com $\mu = 2$

Amostra	$\bar{y}$	Intervalo de 95%	Contém μ?
1	1,98	(1,36;2,60)	Sim
2	1,43	(0,81;2,05)	Sim
3	1,65	(1,03;2,27)	Sim
4	1,88	(1,26;2,50)	Sim
5	2,34	(1,72;2,96)	Sim
6	2,58	(1,96;3,20)	Sim

(Continua...)

Tabela C.2 (continuação)

Intervalos de Confiança Simulados para uma Distribuição Normal(μ,1) com μ = 2

Amostra	$\bar{y}$	Intervalo de 95%	Contém μ?
7	1,58	(0,96;2;20)	Sim
8	2,23	(1,61;2,85)	Sim
9	1,96	(1,34;2,58)	Sim
10	2,11	(1,49;2,73)	Sim
11	2,15	(1,53;2,77)	Sim
12	1,93	(1,31;2,55)	Sim
13	2,02	(1,40;2,64)	Sim
14	2,10	(1,48;2,72)	Sim
15	2,18	(1,56;2,80)	Sim
16	2,10	(1,48;2,72)	Sim
17	1,94	(1,32;2,56)	Sim
18	2,21	(1,59;2,83)	Sim
19	1,16	(0,54;1,78)	Não
20	1,75	(1,13;2,37)	Sim

Intervalos de Confiança para a Média de uma População Normalmente Distribuída

O intervalo de confiança derivado na equação (C.18) ajuda a ilustrar como construir e interpretar intervalos de confiança. Na prática, a equação (C.18) não é muito útil para a média de uma população normal porque ela assume que a variância é conhecida e igual à unidade.

É fácil estender (C.18) para o caso no qual o desvio-padrão σ é conhecido e pode ser qualquer valor: o intervalo de confiança de 95% será

$$[\bar{y} - 1{,}96\sigma/\sqrt{n}, \bar{y} + 1{,}96\sigma\sqrt{n}]. \tag{C.20}$$

Portanto, desde que σ seja conhecido, um intervalo de confiança para μ será prontamente construído. Para possibilitar o uso de σ desconhecido, precisaremos usar uma estimativa. Seja

$$s = \left(\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right)^{1/2} \tag{C.21}$$

o desvio-padrão amostral. Então, obtemos um intervalo de confiança, que dependerá inteiramente dos dados observados, pela substituição de σ na equação (C.20) por sua estimativa, s. Infelizmente, isso não preservará o nível de confiança de 95%, porque s depende da amostra especificada. Em outras palavras, o intervalo aleatório $[\overline{Y} \pm 1{,}96(S/\sqrt{n})]$ não mais conterá μ com probabilidade 0,95, pois a constante σ foi substituída pela variável aleatória S.

Como devemos proceder? Em vez de usar a distribuição normal padrão, devemos recorrer à distribuição t. A distribuição t surge do fato de que

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \tag{C.22}$$

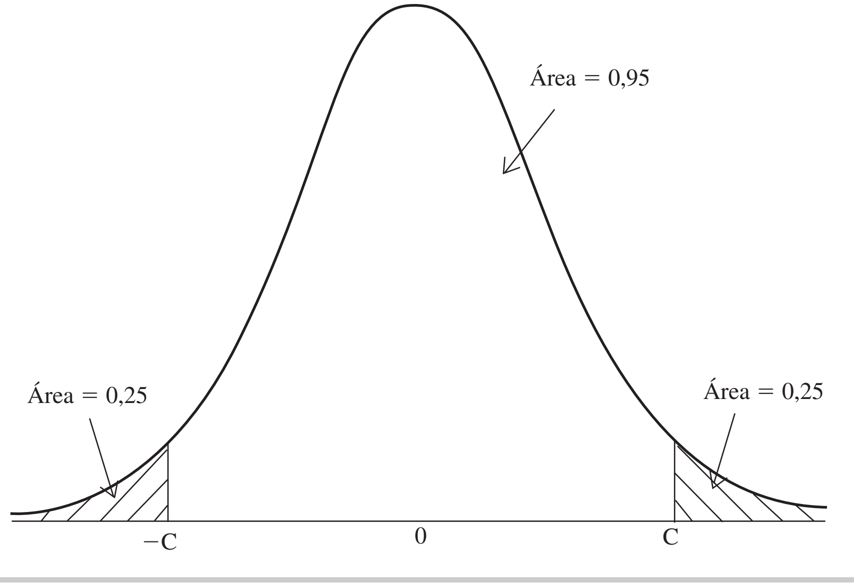
onde $\overline{Y}$ é a média amostral e S é o desvio-padrão amostral da amostra aleatória $\{Y_1, ..., Y_n\}$. Não provaremos (C.22); uma prova cuidadosa pode ser encontrada em diversos lugares [por exemplo, Larsen e Marx (1988, Capítulo 7)].

Para construir um intervalo de confiança, seja c o 97,5º percentil na distribuição t_{n-1}. Em outras palavras, c é o valor tal que 95% da área em t_{n-1} estará entre $-c$ e c: $P(-c < t_{n-1} < c) = 0{,}95$. (O valor de c depende dos graus de liberdade $n - 1$, mas não tornamos isso explícito.) A escolha da c está ilustrada na Figura C.4. Uma vez c tenha sido escolhido de maneira apropriada, o intervalo aleatório $[\overline{Y} - c \cdot S/\sqrt{n}, \overline{Y} + c \cdot S/\sqrt{n}]$ conterá μ com probabilidade 0,95. Para uma determinada amostra, o intervalo de confiança de 95% será calculado como

$$[\bar{y} - c \cdot s/\sqrt{n}, \bar{y} + c \cdot s/\sqrt{n}]. \tag{C.23}$$

Figura C.4

O 97,5º percentil, c, em uma distribuição t.



Os valores de c para vários graus de liberdade podem ser obtidos da Tabela G.2 no Apêndice G. Por exemplo, se $n = 20$, de forma que gl seja $n - 1 = 19$, então, $c = 2{,}093$. Assim, o intervalo de confiança de 95% será $[\bar{y} \pm 2{,}093(s/\sqrt{20})]$, onde $\bar{y}$ e s são os valores obtidos da amostra. Mesmo se $s = \sigma$ (o que será muito pouco provável), o intervalo de confiança em (C.23) será mais amplo que o de (C.20), pois $c > 1{,}96$. Para poucos graus de liberdade, (C.23) será muito mais amplo.

De forma mais geral, seja c_α o percentil $100(1 - \alpha)$ na distribuição t_{n-1}. Então, um intervalo de confiança de $100(1 - \alpha)\%$ será obtido como

$$[\bar{y} - c_{\alpha/2}s/\sqrt{n}, \bar{y} + c_{\alpha/2}s/\sqrt{n}]. \quad \textbf{(C.24)}$$

A obtenção de $c_{\alpha/2}$ exige que se escolha α e o conhecimento dos graus de liberdade $n - 1$; depois, a tabela G.2 poderá ser usada. Na maior parte do tempo, nos concentraremos em intervalos de confiança de 95%.

Existe uma maneira simples de se lembrar como construir um intervalo de confiança para a média de uma distribuição normal. Recorde que $\mathrm{dp}(\bar{Y}) = \sigma/\sqrt{n}$. Assim, $s/\sqrt{n}$ é a estimativa por ponto de $\mathrm{dp}(\bar{Y})$. A variável aleatória associada, $S/\sqrt{n}$, algumas vezes é chamada de **erro-padrão** de $\bar{Y}$. Como o que aparece nas fórmulas é a estimativa por ponto $s/\sqrt{n}$, definimos o erro-padrão de $\bar{y}$ como $\mathrm{ep}(\bar{y}) = s/\sqrt{n}$. Então, (C.24) pode ser escrita abreviadamente como

$$[\bar{y} \pm c_{\alpha/2} \cdot \mathrm{ep}(\bar{y})]. \quad \textbf{(C.25)}$$

Essa equação mostra porque a noção do erro-padrão de uma estimativa desempenha um papel importante em econometria.

EXEMPLO C.2

(Efeitos dos Subsídios de Treinamento de Pessoal sobre a Produtividade dos Trabalhadores)

Holzer, Block, Cheatham e Knott (1993) estudaram os efeitos dos subsídios de treinamento de pessoal sobre a produtividade dos trabalhadores, coletando informações sobre "taxas de rejeição" de uma amostra de empresas industriais de Michigan que haviam recebido subsídios de treinamento de pessoal em 1988. A Tabela C.3 relaciona as taxas de rejeição — medidas como o número de itens, de cada 100 produzidos, que não estavam em condições de uso e, portanto, seriam rejeitados — para 20 empresas. Cada uma dessas empresas recebeu subsídios de treinamento de pessoal em 1988; não houve subsídios em 1987. Estamos interessados em construir um intervalo de confiança para a mudança na taxa de rejeição de 1987 para 1988, para a população de todas as empresas industriais que poderiam ter recebido subsídios.

Assumimos que a mudança nas taxas de rejeição tem uma distribuição normal. Como $n = 20$, um intervalo de confiança de 95% da mudança média nas taxas de rejeição μ será $[\bar{y} \pm 2{,}093 \cdot \mathrm{ep}(\bar{y})]$, onde $\mathrm{ep}(\bar{y}) = s/\sqrt{n}$. O valor 2,093 é o 97,5º percentil em uma distribuição t_{19}. Para os valores amostrais específicos, $\bar{y} = -1{,}15$ e $\mathrm{ep}(\bar{y}) = 0{,}54$ (arredondados para duas casas decimais), e, assim, o intervalo de confiança de 95% será $[-2{,}28, -0{,}02]$. O valor zero foi excluído desse intervalo, de modo que concluímos que, com confiança de 95%, a mudança média nas taxas de rejeição na população não será zero.

EXEMPLO C.2 (continuação)

Tabela C.3

Taxas de Rejeição de 20 Empresas Industriais de Michigan

Empresa	1987	1988	Alteração
1	10	3	−7
2	1	1	0
3	6	5	−1
4	0,45	0,5	0,05
5	1,25	1,54	0,29
6	1,3	1,5	0,2
7	1,06	0,8	−0,26
8	3	2	−1
9	8,18	0,67	−7,51
10	1,67	1,17	−0,5
11	0,98	0,51	−0,47
12	1	0,5	−0,5
13	0,45	0,61	0,16
14	5,03	6,7	1,67
15	8	4	−4
16	9	7	−2
17	18	19	1
18	0,28	0,2	−0,08
19	7	5	−2
20	3,97	3,83	−0,14
Média	4,38	3,23	−1,15

Neste ponto, o Exemplo C.2 é bastante ilustrativo, pois ele tem algumas imperfeições potencialmente sérias como uma análise econométrica. De forma mais importante, ele assume que qualquer redução sistemática nas taxas de rejeição deve-se aos subsídios de treinamento de pessoal. Entretanto, muita coisa pode acontecer no decurso do ano para alterar a produtividade dos trabalhadores. A partir dessa análise, não temos meios de saber se a queda na média das taxas de rejeição é atribuível aos subsídios de treinamento ou se, pelo menos parcialmente, outra causa externa foi a responsável.

Uma Regra Prática Simples para um Intervalo de Confiança de 95%

O intervalo de confiança em (C.25) pode ser calculado para qualquer tamanho de amostra e qualquer nível de confiança. Como vimos na seção B.4, a distribuição t se aproxima da distribuição normal padrão conforme os graus de liberdade aumentam. Particularmente, para $\alpha = 0{,}05$, $c_{\alpha/2} \rightarrow 1{,}96$ quando $n \rightarrow \infty$, embora $c_{\alpha/2}$ seja sempre maior que 1,96 para cada n. Uma *regra prática* para um intervalo de confiança aproximado de 95% é

$$[\bar{y} \pm 2 \cdot \mathrm{ep}(\bar{y})]. \qquad \textbf{(C.26)}$$

Em outras palavras, obtemos $\bar{y}$ e seu erro-padrão e depois calculamos $\bar{y}$ mais e menos duas vezes seu erro-padrão para obter o intervalo de confiança. Isso é um pouco amplo demais para n muito grande, e estreito demais para n pequeno. Como podemos ver pelo exemplo C.2, mesmo para um n tão pequeno como 20, (C.26) está muito próximo do intervalo de confiança de 95% da média de uma distribuição normal. Isso significa que podemos chegar muito próximos de um intervalo de confiança de 95% sem termos de recorrer às tabelas t.

Intervalos de Confiança Assimptóticos para Populações Não-Normais

Em algumas aplicações, a população é claramente não-normal. Um caso destacado é a distribuição de Bernoulli, na qual a variável aleatória assume somente os valores zero e um. Em outros casos, a população não-normal não tem qualquer distribuição padrão. Isso não tem importância, desde que o tamanho da amostra seja suficientemente grande para que o teorema do limite central produza uma boa aproximação da distribuição da média amostral $\bar{Y}$. Para n grande, um intervalo de confiança de 95% *aproximado* será

$$[\bar{y} \pm 1{,}96 \cdot \mathrm{ep}(\bar{y})], \qquad \textbf{(C.27)}$$

onde o valor 1,96 é o 97,5º percentil na distribuição normal padrão. Mecanicamente, calcular um intervalo de confiança aproximado não difere do caso normal. Uma pequena diferença é o fato de o número que multiplica o erro padrão vir da distribuição normal padrão, em vez da distribuição t, pois estamos usando um tratamento assimptótico. Como a distribuição t se aproxima da normal padrão à medida que os *gl* aumentam, a equação (C.25) também é perfeitamente legítima como um intervalo aproximado de 95%; alguns preferem essa equação a (C.27), pois a primeira é exata para populações normais.

EXEMPLO C.3

(Discriminação Racial na Contratação de Trabalhadores)

O Urban Institute conduziu um estudo em 1988, em Washington, D.C., para examinar a extensão da discriminação racial na contratação de trabalhadores. Cinco duplas de pessoas foram entrevistadas para várias ofertas de emprego. Em cada dupla, uma pessoa era negra, e a outra branca. Todos portavam currículos indicando que tinham virtualmente os mesmos graus de experiência, educação e outros fatores que determinavam a qualificação para os cargos. A idéia era tornar os indivíduos tão semelhantes quanto possível, com exceção da raça. Cada pessoa da mesma dupla se candidatou ao mesmo emprego, e os pesquisadores registraram quem recebeu uma oferta de emprego. Esse é um exemplo de uma análise de pares comparados, na qual cada observação consiste

EXEMPLO C.3 (continuação)

de dados sobre duas pessoas (ou duas empresas, duas cidades etc.) que são tidas como semelhantes em muitos aspectos, mas diferentes em uma característica importante.

Seja θ_N a probabilidade de que a pessoa negra receba uma oferta de emprego e θ_B seja a probabilidade de que a pessoa branca receba a oferta. Estamos basicamente interessados na diferença $\theta_N - \theta_B$. Seja N_i uma variável de Bernoulli igual a um se a pessoa negra conseguir uma oferta de emprego do empregador i, e zero, caso contrário. Semelhantemente, $B_i = 1$ se a pessoa branca conseguir uma oferta de emprego do empregador i, e zero, caso contrário. Agrupando as cinco duplas de pessoas, houve um total de $n = 241$ observações (pares de entrevistas com os candidatos). Estimadores não-viesados de θ_N e θ_B são $\overline{N}$ e $\overline{B}$, as frações de entrevistas para as quais foram oferecidas propostas de emprego aos negros e brancos, respectivamente.

Para colocar tudo isso em uma estrutura para calcular um intervalo de confiança de uma média populacional, defina uma nova variável $Y_i = N_i - B_i$. Agora, Y_i pode assumir três valores: -1 se a pessoa negra não recebeu a proposta de emprego, mas a pessoa branca recebeu, 0 se ambas as pessoas conseguiram ou não o emprego, e 1 se a pessoa negra conseguiu o emprego e a pessoa branca não. Então, $\mu \equiv \mathrm{E}(Y_i) = \mathrm{E}(N_i) - \mathrm{E}(B_i) = \theta_N - \theta_B$.

A distribuição de Y_i certamente não é normal — ela é discreta e assume somente três valores. No entanto, um intervalo de confiança aproximado de $\theta_N - \theta_B$ pode ser obtido usando métodos de amostras grandes.

Usando os 241 pontos de dados observados, $\overline{b} = 0{,}224$ e $\overline{w} = 0{,}357$, de modo que $\overline{y} = 0{,}224 - 0{,}357 = -0{,}133$. Assim, 22,4% dos candidatos negros receberam oferta de emprego, enquanto a oferta de emprego foi oferecida a 35,7% dos brancos. Isso é evidência *prima facie* de discriminação contra os negros, mas podemos descobrir muito mais calculando um intervalo de confiança para μ. Para calcular um intervalo de confiança aproximado de 95%, precisamos do desvio-padrão da amostra. Obtemos $s = 0{,}482$ [usando a equação (C.21)]. Usando (C.27), obteremos um IC de 95% de $\mu = \theta_N - \theta_B$ como $-0{,}133 \pm 1{,}96(0{,}482/\sqrt{241}) = -0{,}133 \pm 0{,}031 = [-0{,}164, -0{,}102]$. O IC de 99% será $-0{,}133 \pm 2{,}58(0{,}482/\sqrt{241}) = [-0{,}213, -0{,}053]$. Naturalmente, esse intervalo contém um leque mais amplo de valores que o IC de 95%. Mas mesmo o IC de 99% não contém o valor zero. Portanto, estamos bastante confiantes que a diferença populacional $\theta_N - \theta_B$ não é zero.

Precisamos fazer um comentário final antes de abandonarmos o tópico de intervalo de confiança. Como o erro-padrão de $\overline{y}$, $\mathrm{ep}(\overline{y}) = s/\sqrt{n}$, se contrai para zero conforme o tamanho da amostra cresce, vemos que — tudo mais mantido igual — um tamanho maior de amostra significa um intervalo de confiança menor. Assim, uma importante vantagem de uma amostra de tamanho grande é que ela resulta em intervalos de confiança menores.

C.6 TESTES DE HIPÓTESES

Até agora, revimos como avaliar estimadores por ponto, e vimos — no caso de uma média populacional — como construir e interpretar intervalos de confiança. Entretanto, algumas vezes a questão na qual estamos interessados tem uma resposta sim ou não bem definida. Eis alguns exemplos: (1) Um programa de treinamento de pessoal efetivamente aumenta a produtividade média dos trabalhadores? (veja o exemplo C.2); (2) os negros são discriminados na contratação de trabalhadores? (veja o exemplo C.3); (3) leis estaduais mais rigorosas contra dirigir embriagado reduzem o número de prisões por esse delito? Os métodos para responder a tais questões, usando uma amostra de dados, são conhecidos como testes de hipóteses.

Fundamentos dos Testes de Hipóteses

Para ilustrar os problemas envolvidos com os testes de hipóteses, considere um exemplo sobre eleições. Suponha que haja dois candidatos em uma eleição, Candidatos A e B. O Candidato A recebeu 42% dos votos populares, enquanto o Candidato B recebeu 58%. Esses números supostamente representam as porcentagens verdadeiras da população votante e serão tratados como tais.

O Candidato A está convencido de que um número maior de pessoas deve ter votado nele, e, assim, ele gostaria de investigar se a eleição foi burlada. Conhecendo um pouco de estatística, esse candidato contrata uma empresa de consultoria para aleatoriamente extrair uma amostra de 100 eleitores para registrar se cada pessoa votou ou não nele. Suponha que, para a amostra coletada, 53 pessoas votaram no Candidato A. Essa estimativa amostral de 53% claramente excede o valor populacional oficial de 42%. O Candidato A deve concluir que a eleição foi realmente uma fraude?

Embora pareça que tenha havido uma menor contagem de votos para o Candidato A, não podemos ter certeza disso. Mesmo se apenas 42% da população tenha votado no Candidato A, é possível que, em uma amostra de 100 eleitores, observemos 53 pessoas que realmente votaram no Candidato A. A questão é: o quanto é *forte* a evidência amostral contra a porcentagem oficial de 42% divulgada?

Uma maneira de proceder é montar um **teste de hipótese**. Seja θ a proporção verdadeira da população que votou no Candidato A. A hipótese de que os resultados divulgados são precisos pode ser definida como

$$\mathrm{H}_0\text{: } \theta = 0{,}42. \tag{C.28}$$

Esse é um exemplo de uma **hipótese nula**. Sempre representamos a hipótese nula por H_0. Nos testes de hipóteses, a hipótese nula tem papel semelhante ao de um réu em julgamento em muitos sistemas judiciais: da mesma forma que se presume que um réu é inocente até que sua culpa seja provada, a hipótese nula é presumida como verdadeira até que os dados sugiram fortemente o contrário. No exemplo em questão, o Candidato A deverá apresentar evidências bastante fortes contra (C.28) para ter direito a uma recontagem dos votos.

A **hipótese alternativa** no exemplo da eleição é que a proporção verdadeira dos votantes no Candidato A na eleição seja maior que 0,42:

$$\mathrm{H}_1\text{: } \theta > 0{,}42. \tag{C.29}$$

Para concluir que H_0 é falsa, e H_1 é verdadeira, precisamos ter evidência "além da dúvida razoável" contra H_0. Quantos votos dos 100 seriam necessários para sentir que a evidência seria fortemente contra H_0? A maioria das pessoas concordaria que a observação de 43 votos em uma amostra de 100 eleitores não seria suficiente para reverter os resultados originais da eleição; tal resultado está bem dentro da variação amostral esperada. Por outro lado, não precisaremos observar 100 votos para o Candidato A para lançarmos dúvidas sobre H_0. Se 53 em 100 é um número suficiente para rejeitar H_0, isso é muito menos claro. A resposta dependerá de como quantificamos a expressão "além da dúvida razoável".

Nos testes de hipóteses, podemos cometer dois tipos de enganos. Primeiro, podemos rejeitar a hipótese nula quando na verdade ela é verdadeira. Esse é o chamado **erro tipo I**. No exemplo das eleições, um erro tipo I ocorrerá se rejeitarmos H_0 quando a proporção verdadeira das pessoas que votaram no Candidato A for de fato 0,42. O segundo tipo de erro é a impossibilidade de rejeitar H_0 quando ela for efetivamente falsa. Esse é o chamado **erro tipo II**. No exemplo das eleições, um erro tipo II ocorrerá se $\theta > 0{,}42$, mas não pudermos rejeitar H_0.

Após termos tomado a decisão de rejeitar ou não a hipótese nula, ou nossa decisão foi correta ou cometemos um erro. Nunca saberemos com certeza se um erro foi cometido. Porém, podemos calcular a *probabilidade* de cometer um erro tipo I ou um erro tipo II. As regras dos testes de hipóteses são construídas para fazer com que a probabilidade de cometer um erro tipo I seja muito pequena. De forma geral, definimos o **nível de significância** (ou simplesmente o *nível*) de um teste como a probabilidade de um erro tipo I; isso é geralmente representado por α. Simbolicamente, temos

$$\alpha = \mathrm{P}(\text{Rejeitar H}_0|\mathrm{H}_0). \tag{C.30}$$

O lado direito da equação é lido como: "A probabilidade de rejeitar H_0 supondo que H_0 seja verdadeira".

Os testes clássicos de hipóteses exigem que inicialmente especifiquemos um nível de significância do teste. Quando especificamos um valor de α, estamos essencialmente quantificando nossa tolerância para um erro tipo I. Valores comuns de α são 0,10, 0,05, e 0,01. Se $\alpha = 0{,}05$, o pesquisador estará querendo falsamente rejeitar H_0 em 5% das vezes, de maneira a detectar desvios em relação a H_0.

Uma vez determinado o nível de significância, então, gostaríamos de minimizar a probabilidade de um erro tipo II. Alternativamente, gostaríamos de maximizar o **poder de um teste** contra todas as alternativas relevantes. O poder de um teste é simplesmente um menos a probabilidade de um erro tipo II. Matematicamente,

$$\pi(\theta) = \mathrm{P}(\text{Rejeitar H}_0|\theta) = 1 - \mathrm{P}(\text{Tipo II}|\theta),$$

onde θ representa o valor efetivo do parâmetro. Naturalmente, gostaríamos que o poder seja igual à unidade sempre que a hipótese nula for falsa. Mas isso é impossível de ser atingido mantendo pequeno o nível de significância. Em vez disso, preferimos que nosso teste maximize o poder para determinado nível de significância.

Testes de Hipóteses sobre a Média em uma População Normal

Para testar a hipótese nula contra uma alternativa, precisamos escolher uma estatística de teste (ou estatística, resumidamente) e um valor crítico. A escolha da estatística e do valor crítico é baseada na conveniência e no desejo de maximizar o poder do teste, dado um nível de significância do teste. Nesta subseção examinaremos como testar hipóteses para a média de uma população normal.

Uma **estatística de teste**, representada por T, é alguma função da amostra aleatória. Quando calculamos a estatística para um determinado resultado, obtemos um resultado da estatística de teste, que denominaremos t.

Dada uma estatística de teste, podemos definir uma regra de rejeição que determine quando H_0 deve ser rejeitada em favor de H_1. Neste texto, todas as regras de rejeição são baseadas na comparação do valor de uma estatística de teste, t, com um **valor crítico**, c. Os valores de t que resultam na rejeição da hipótese nula são coletivamente conhecidos como **região de rejeição**. Para determinarmos o valor crítico, primeiro devemos decidir sobre um nível de significância do teste. Em seguida, dado α, o valor crítico associado com α é determinado pela distribuição de T, *assumindo* que H_0 seja verdadeira. Escreveremos esse valor crítico como c, omitindo o fato de que ele depende de α.

Testar hipóteses sobre a média μ de uma população Normal(μ,σ^2) é simples. A hipótese nula é definida como

$$H_0: \mu = \mu_0, \tag{C.31}$$

onde μ_0 é um valor que especificamos. Na maioria das aplicações, $\mu_0 = 0$, mas o caso generalizado não é mais difícil do que isso.

A regra de rejeição que escolheremos dependerá da natureza da hipótese alternativa. As três alternativas de interesse são

$$H_1: \mu > \mu_0, \tag{C.32}$$

$$H_1: \mu < \mu_0, \tag{C.33}$$

e

$$H_1: \mu \neq \mu_0. \tag{C.34}$$

A equação (C.32) produz uma **alternativa unilateral**, como também (C.33). Quando a hipótese alternativa for (C.32), a hipótese nula será efetivamente H_0: $\mu \leq \mu_0$, já que somente rejeitaremos H_0 quando $\mu > \mu_0$. Isso será apropriado quando estivermos interessados no valor de μ somente quando μ for pelo menos tão grande quanto μ_0. A equação (C.34) é uma **alternativa bilateral**. Ela será apropriada quando estivermos interessados em qualquer desvio da hipótese nula.

Considere primeiro a alternativa (C.32). Intuitivamente, deveríamos rejeitar H_0 em favor de H_1 quando o valor da média amostral, $\bar{y}$, fosse "suficientemente" maior que μ_0. Mas como devemos determinar quando $\bar{y}$ é grande o suficiente para que H_0 seja rejeitada ao nível de significância escolhido? Isso requer que se conheça a probabilidade de rejeitar a hipótese nula quando ela for verdadeira. Em vez de trabalhar diretamente com $\bar{y}$, usamos sua versão padronizada, na qual σ é substituído pelo desvio-padrão amostral, s:

$$t = \sqrt{n}(\bar{y} - \mu_0)/s = (\bar{y} - \mu_0)/\text{ep}(\bar{y}), \tag{C.35}$$

onde $\text{ep}(\bar{y}) = s/\sqrt{n}$ é o erro-padrão de $\bar{y}$. Dada a amostra de dados, é fácil obter t. A razão pela qual trabalhamos com t é que, de acordo com a hipótese nula, a variável aleatória

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S$$

tem uma distribuição t_{n-1}. Agora, suponha que tenhamos nos fixado em um nível de significância de 5%. Então, o valor crítico c será determinado de forma que $P(T > c|H_0) = 0{,}05$; ou seja, a probabilidade de um erro tipo I é de 5%. Uma vez encontrado c, a regra de rejeição será

$$t > c, \tag{C.36}$$

onde c é o percentil $100(1 - \alpha)$ em uma distribuição t_{n-1}; em forma de porcentagem, o nível de significância é $100 \cdot \alpha\%$. Esse é um exemplo de um **teste monocaudal**, pois a região de rejeição está em uma extremidade da distribuição t. Para um nível de significância de 5%, c será o 95º percentil na distribuição t_{n-1}; isso está ilustrado na Figura C.5. Um nível diferente de significância leva a um valor crítico diferente.

A estatística na equação (C.35) muitas vezes é chamada de **estatística *t*** para testar H_0: $\mu = \mu_0$. A estatística t mede a distância de $\bar{y}$ a μ_0 em relação ao erro-padrão de $\bar{y}$, ep($\bar{y}$).

EXEMPLO C.4

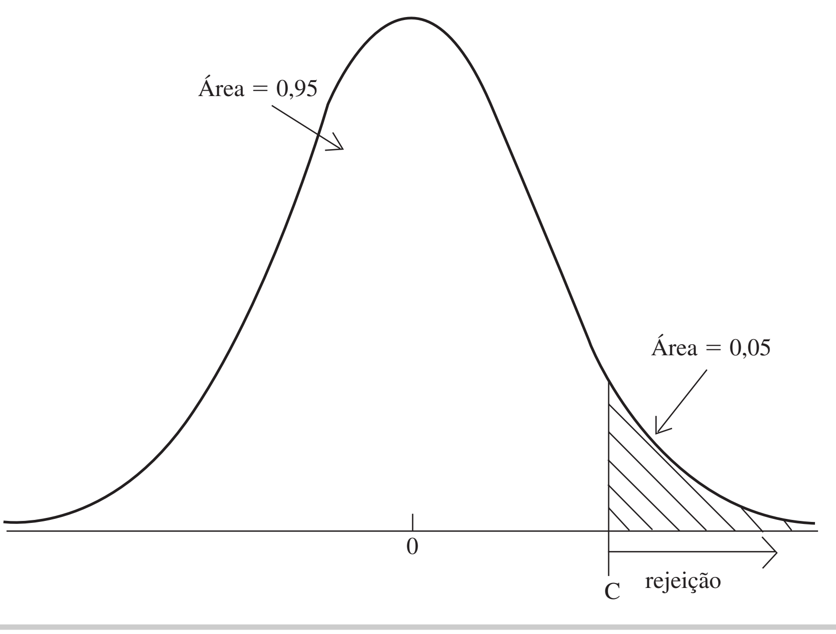
(Efeito das Zonas Industriais sobre os Investimentos Empresariais)

Na população de cidades onde foram criadas zonas industriais em determinado estado [veja o caso de Indiana em Papke (1994)], seja Y a mudança percentual nos investimentos do ano anterior ao ano posterior em que uma cidade se tornou uma zona industrial. Assuma que Y tem uma distribuição Normal(μ,σ^2). A hipótese nula de que zonas industriais não têm efeito nos investimentos é H_0: $\mu = 0$; a alternativa de que elas têm um efeito positivo é H_1: $\mu > 0$. (Assumimos que elas não têm um efeito negativo). Suponha que queremos testar H_0 ao nível de 5%. O teste estatístico nesse caso será

$$t = \frac{\bar{y}}{s/\sqrt{n}} = \frac{\bar{y}}{\text{ep}(\bar{y})}. \tag{C.37}$$

Figura C.5

Região de rejeição para um teste ao nível de significância de 5% contra a alternativa unilateral $\mu > \mu_0$.



EXEMPLO C.4 (continuação)

Suponha que temos uma amostra de 36 cidades onde foram criadas zonas industriais. Nesse caso, o valor crítico será $c = 1{,}69$ (veja Tabela G.2), e rejeitamos H_0 em favor de H_1 se $t > 1{,}69$. Suponha que a amostra produza $\bar{y} = 8{,}2$ e $s = 23{,}9$. Então, $t \approx 2{,}06$ e H_0 será, portanto, rejeitada ao nível de 5%. Assim, concluímos que, ao nível de significância de 5%, as zonas industriais têm um efeito sobre o investimento médio. O valor crítico de 1% será 2,44, e, portanto, H_0 não será rejeitada ao nível de 1%. A mesma limitação do Exemplo C.2 é válida neste caso: não controlamos os outros fatores que possam afetar o investimento nas cidades ao longo do tempo, e, portanto, não podemos afirmar que o efeito seja causal.

A regra de rejeição é semelhante para a alternativa unilateral (C.33). Um teste com nível de significância $100 \cdot \alpha\%$ rejeitará H_0 contra (C.33) sempre que

$$t < -c; \qquad \textbf{(C.38)}$$

em outras palavras, estamos procurando por valores negativos da estatística t — o que implica que $\bar{y} < \mu_0$ — que estejam suficientemente distantes de zero para rejeitar H_0.

Para alternativas bilaterais, devemos ter cuidado na escolha do valor crítico, de forma que o nível de significância do teste ainda seja α. Se H_1 for dada por H_1: $\mu \neq \mu_0$, então, rejeitaremos H_0 se $\bar{y}$ estiver distante da μ_0 em *valor absoluto*: um $\bar{y}$ muito maior ou muito menor que μ_0 fornece evidência contra H_0 em favor de H_1. Um nível $100 \cdot \alpha\%$ para o teste é obtido pela regra de rejeição

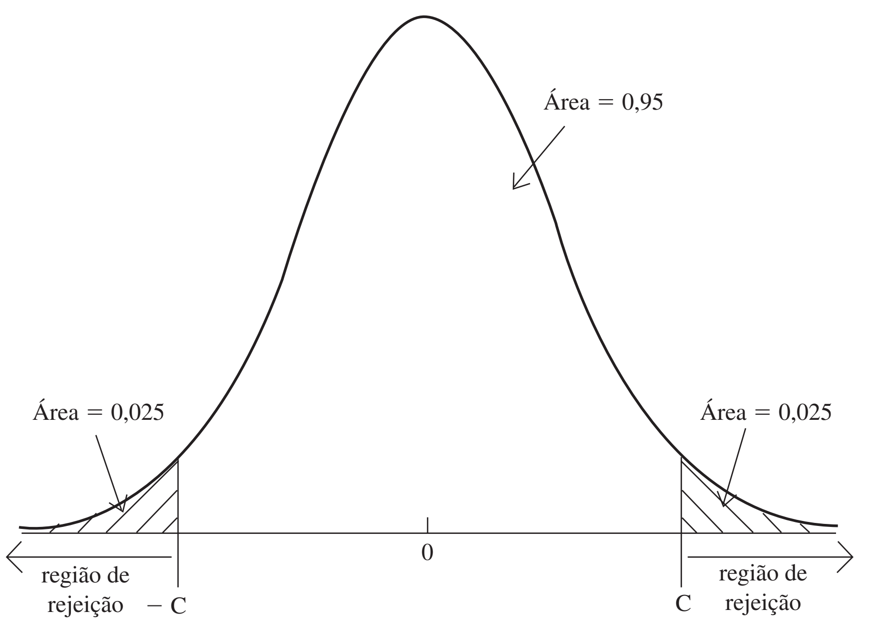
$$|t| > c, \qquad \textbf{(C.39)}$$

onde $|t|$ é o valor absoluto da estatística t em (C.35). Isso produz um **teste bicaudal**. Agora precisamos ser cuidadosos na escolha do valor crítico: c é o $100(1 - \alpha/2)$ percentil na distribuição t_{n-1}. Por exemplo, se $\alpha = 0{,}05$, então, o valor crítico será o 97,5º percentil na distribuição t_{n-1}. Isso garante que H_0 será rejeitada em somente 5% das vezes quando ela for verdadeira (veja a Figura C.6). Por exemplo, se $n = 22$, então, o valor crítico c será 2,08, o 97,5º percentil em uma distribuição t_{21} (veja a Tabela G.2). O valor absoluto da estatística t deve exceder 2,08 para rejeitar H_0 contra H_1 ao nível de 5%.

É importante conhecer a linguagem apropriada dos testes de hipóteses. Algumas vezes, a frase apropriada "não podemos rejeitar H_0 em favor de H_1 ao nível de significância de 5%" é substituída por "aceitamos H_0 ao nível de significância de 5%". A última construção é incorreta. Com o mesmo conjunto de dados, geralmente existem muitas hipóteses que não podem ser rejeitadas. No exemplo anterior das eleições, seria logicamente inconsistente dizer que H_0: $\theta = 0{,}42$ e H_0: $\theta = 0{,}43$ são ambas "aceitas", pois somente uma delas pode ser verdadeira. Entretanto, é completamente possível que nenhuma dessas hipóteses seja rejeitada. Por essa razão, sempre dizemos "não ser possível rejeitar H_0" em vez de "aceitar H_0".

Figura C.6

Região de rejeição para um teste ao nível de significância de 5% contra a alternativa bilateral $H_1: \mu \neq \mu_0$.



Testes Assimptóticos para Populações Não-Normais

Se o tamanho da amostra for suficientemente grande para invocar o teorema do limite central (veja Seção C.3), a mecânica dos testes de hipóteses de médias populacionais será a mesma, seja ou não normal a distribuição amostral. A justificação teórica vem do fato que, de acordo com a hipótese nula,

$$T = \sqrt{n}(\overline{Y} - \mu_0)/S \overset{a}{\sim} \text{Normal}(0,1).$$

Portanto, com n grande, podemos comparar a estatística t em (C.35) com os valores críticos de uma distribuição normal padrão. Como a distribuição t_{n-1} converge para a distribuição normal padrão à medida que n vai ficando maior, os valores críticos de t e da distribuição normal padrão ficarão muito próximos com n extremamente grande. Como a teoria assimptótica é baseada em n crescendo sem limites, ela não pode nos informar quais valores críticos são os melhores, se da normal padrão ou de t. Para valores moderados de n, digamos entre 30 e 60, é tradicional usar a distribuição t, pois sabemos que ela é correta para populações normais. Para $n > 120$, a escolha entre as distribuições t e normal padrão é largamente irrelevante, porque os valores críticos são praticamente os mesmos.

Como os valores críticos escolhidos usando a distribuição normal padrão ou a distribuição t serão somente aproximadamente válidos para populações não-normais, nosso níveis de significância escolhidos serão também apenas aproximados; assim, para populações não-normais, nossos níveis de significância serão realmente assimptóticos. Dessa forma, se escolhermos um nível de significância de 5%, mas nossa população for não-normal, então, o nível de significância efetivo será maior ou menor que 5% (e não teremos como saber qual será o caso). Quando o tamanho da amostra é grande, o nível de significância efetivo estará muito próximo de 5%. De modo prático, a distinção não é importante, e, portanto, não mais usaremos a qualificação "assimptótico".

EXEMPLO C.5

(Discriminação Racial na Contratação de Trabalhadores)

No estudo do Urban Institute sobre a discriminação racial na contratação de trabalhadores (veja o exemplo C.3), estamos essencialmente interessados em testar H_0: $\mu = 0$ contra H_1: $\mu < 0$, onde $\mu = \theta_N - \theta_B$ é a diferença em probabilidades de que negros e brancos recebam ofertas de emprego. Recorde que μ é a média populacional da variável $Y = N - B$, onde N e B são indicadores binários. Usando as $n = 241$ comparações de duplas, obtivemos $\bar{y} = -0{,}133$ e ep$(\bar{y}) = 0{,}482/\sqrt{241} \approx 0{,}031$. A estatística t para testar H_0: $\mu = 0$ é $t = -0{,}133/0{,}031 \approx -4{,}29$. Você se lembrará do Apêndice B que a distribuição normal padrão é, para propósitos práticos, indistinguível da distribuição t com 240 graus de liberdade. O valor $-4{,}29$ está tão distante da extremidade esquerda da distribuição que rejeitamos H_0 a qualquer nível razoável de significância. Aliás, o valor crítico (do teste unilateral) de 0,005 (metade de um porcento) está em torno de $-2{,}58$. Um valor t de $-4{,}29$ é evidência muito forte contra H_0 em favor de H_1. Portanto, concluímos que existe discriminação na contratação de trabalhadores.

Cálculo e Uso de *p*-Valores

O requisito tradicional de se escolher um nível de significância antes do tempo quer dizer que diferentes pesquisadores, usando os mesmos dados e o mesmo procedimento para testar a mesma hipótese poderiam terminar com conclusões diferentes. A divulgação do nível de significância no qual estamos fazendo nosso teste resolve este problema até certo ponto, mas não elimina completamente o problema.

Para fornecer mais informação, podemos fazer a seguinte pergunta: qual é o maior nível de significância no qual poderíamos conduzir nosso teste e ainda não conseguir rejeitar a hipótese nula? Esse valor é conhecido como o ***p*-valor** de um teste (algumas vezes chamado de *prob-valor*). Comparado com a escolha de um nível de significância antes do tempo e a obtenção de um valor crítico, calcular um p-valor é um pouco mais difícil. Entretanto, com o advento da computação rápida e barata, p-valores são agora razoavelmente fáceis de serem obtidos.

Como ilustração, considere o problema de testar H_0: $\mu = 0$ em uma população Normal(μ,σ^2). Nossa estatística de teste nesse caso será $T = \sqrt{n} \cdot \bar{Y}/S$, e assumimos que n é grande o suficiente para tratar T como uma distribuição normal padrão sob H_0. Suponha que o valor observado de T para nossa amostra seja $t = 1{,}52$. (Observe como pulamos o passo da escolha de um nível de significância.) Agora que já vimos o valor t, podemos encontrar o maior nível de significância no qual não conseguiríamos rejeitar H_0. Esse será o nível de significância associado ao uso de t como nosso valor crítico. Como nosso teste estatístico T tem uma distribuição normal padrão sob H_0, teremos

$$p\text{-valor} = P(T > 1{,}52|H_0) = 1 - \Phi(1{,}52) = 0{,}065, \quad \textbf{(C.40)}$$

onde $\Phi(\cdot)$ representa a fdc normal padrão. Em outras palavras, o p-valor neste exemplo é simplesmente a área à direita de 1,52, o valor observado da estatística de teste, em uma distribuição normal padrão. Veja a Figura C.7 a título de ilustração.

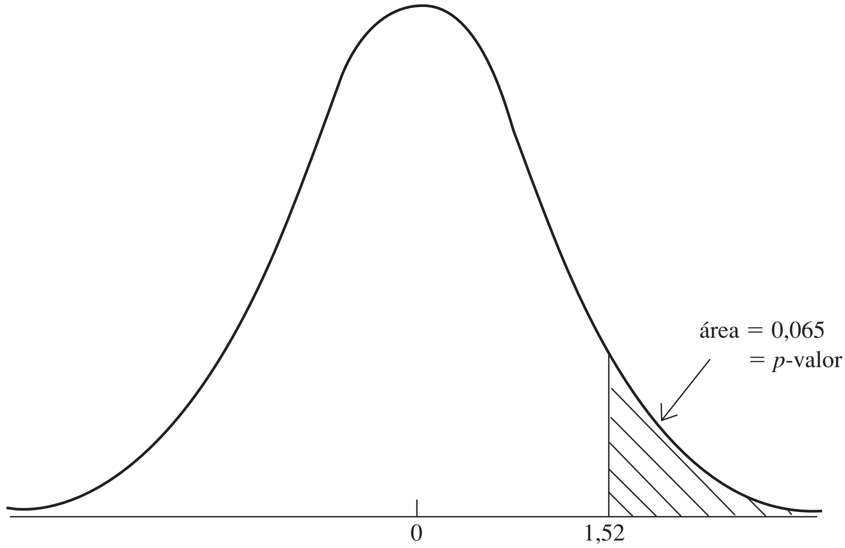
Como p-valor $= 0{,}065$, o maior nível de significância no qual poderemos conduzir este teste e não conseguir rejeitar H_0 será de 6,5%. Se executarmos o teste a um nível abaixo de 6,5% (como, por exemplo, 5%), não rejeitaremos H_0. Se executarmos o teste a um nível maior que 6,5% (como, por exemplo,10%), rejeitaremos H_0. Com o p-valor à mão, poderemos conduzir o teste em qualquer nível.

O p-valor nesse exemplo tem outra interpretação útil: ele é a probabilidade de que observemos um valor de T tão grande quanto 1,52 quando a hipótese nula for verdadeira. Se a hipótese nula for efetivamente verdadeira, observaremos um valor de T tão grande quanto 1,52 devido ao acaso de somente em 6,5% das vezes. Se isso será suficientemente pequeno para rejeitar H_0 dependerá de nossa tolerância de um erro tipo I. O p-valor tem uma interpretação semelhante em todos os outros casos, como veremos.

De forma geral, p-valores pequenos são evidência *contra* H_0, já que eles indicam que o resultado dos dados ocorrerá com pequena probabilidade se H_0 for verdadeira. No exemplo anterior, se t tivesse um valor maior, digamos $t = 2{,}85$, então, o p-valor teria sido $1 - \Phi(2{,}85) \approx 0{,}002$. Isso significaria que, se a hipótese nula fosse verdadeira, observaríamos um valor de T tão grande quanto 2,85 com probabilidade 0,002. Como interpretamos isso? Ou obtivemos uma amostra bastante incomum ou a hipótese nula é falsa. A menos que tivéssemos uma tolerância muito pequena para um erro tipo I, teríamos rejeitado a hipótese nula. Por outro lado, um p-valor grande é uma fraca evidência contra H_0. Se tivéssemos obtido $t = 0{,}47$ no exemplo anterior, então, p-valor $= 1 - \Phi(0{,}47) = 0{,}32$. A observação de um valor de T maior que 0,47 aconteceria com probabilidade 0,32, mesmo quando H_0 fosse verdadeira; isso seria suficientemente grande para que não houvesse dúvida quanto a H_0, a menos que tivéssemos uma tolerância muito alta para o erro tipo I.

Figura C.7

O p-valor quando $t = 1{,}52$ para a alternativa unilateral $\mu > \mu_0$.



Para testarmos hipóteses sobre uma média populacional usando a distribuição t, precisamos de tabelas detalhadas para computar p-valores. A Tabela G.2 somente nos possibilita colocar delimitadores nos p-valores. Felizmente, muitos programas estatísticos e econométricos agora computam p-valores de forma rotineira, e eles também fornecem cálculos de fdcs para a distribuição t e outras usadas na computação dos p-valores.

EXEMPLO C.6

(Efeitos dos Subsídios de Treinamento de Pessoal sobre a Produtividade dos Trabalhadores)

Considere novamente os dados de Holzer et al. (1993) no Exemplo C.2. De uma perspectiva de planejamento, há duas questões de interesse. Primeiro, qual será nossa melhor estimativa da alteração da média nas taxas de rejeição, μ? Já obtivemos isso para a amostra de 20 empresas listadas na Tabela C.3: a média amostral da alteração nas taxas de rejeição foi de $-1{,}15$. Em relação à taxa média de rejeição inicial em 1987, isso representa uma queda na taxa de rejeição de cerca de 26,3% ($-1{,}15/4{,}38 \approx -0{,}263$) que é um efeito nada desprezível.

Também gostaríamos de saber se a amostra fornece forte evidência de um efeito na população de empresas industriais que poderiam ter recebido subsídios de treinamento de pessoal. A hipótese nula é H_0: $\mu = 0$, que foi testada contra H_1: $\mu < 0$, onde μ é a alteração média nas taxas de rejeição. Sob a hipótese nula, os subsídios de treinamento de pessoal não têm efeito sobre as taxas de rejeição. A hipótese alternativa estabelece que existe um efeito. Não nos importamos com a alternativa $\mu > 0$, de modo que a hipótese nula será efetivamente H_0: $\mu \geq 0$.

Como $\bar{y} = -1{,}15$ e ep($\bar{y}$) $= 0{,}54$, $t = -1{,}15/0{,}54 = -2{,}13$. Esse valor está abaixo do valor crítico de $-1{,}73$ ao nível de 5% (de uma distribuição t_{19}), mas acima do valor crítico de $-2{,}54$. O p-valor nesse caso será computado da seguinte forma

$$p\text{-valor} = \mathrm{P}(T_{19} < -2{,}13), \tag{C.41}$$

onde T_{19} representa uma variável aleatória com distribuição t, com 19 graus de liberdade. A desigualdade é o oposto de (C.40), pois a hipótese alternativa tem a forma de (C.33). A probabilidade em (C.41) é a área à esquerda de $-2{,}13$ em uma distribuição t_{19} (veja a Figura C.8).

Usando a Tabela G.2, o máximo que poderemos dizer é que o p-valor estará entre 0,025 e 0,01, mas ele estará mais próximo de 0,025 (já que o 97,5º percentil é cerca de 2,09). Utilizando um programa estatístico, como o Stata, podemos computar o p-valor exato. Ele será 0,023, que é evidência razoável contra H_0. Isso certamente é evidência suficiente para rejeitar a hipótese nula de que os subsídios de treinamento de pessoal não têm qualquer efeito ao nível de significância de 2,5% (e, portanto, ao nível de 5%).

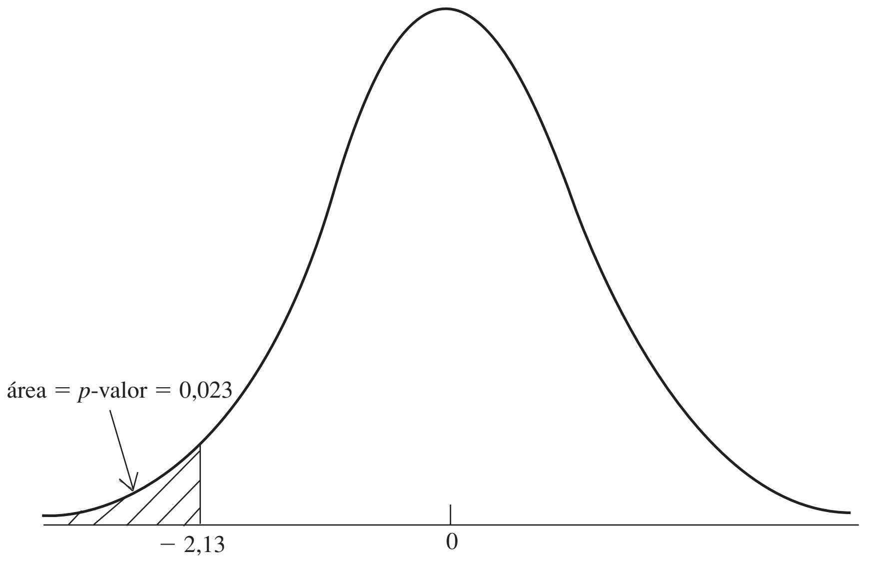
O cálculo de um p-valor para um teste bilateral é semelhante, mas devemos considerar a natureza bilateral da regra de rejeição. Para testes t sobre médias populacionais, o p-valor é computado como

$$\mathrm{P}(|T_{n-1}| > |t|) = 2\mathrm{P}(T_{n-1} > |t|), \tag{C.42}$$

onde t é o valor da estatística do teste e T_{n-1} é uma variável aleatória t. (Para n grande, substitua T_{n-1} por uma variável aleatória normal padrão). Assim, compute o valor absoluto da estatística t, encontre a área à direita desse valor em uma distribuição t_{n-1}, e multiplique a área por dois.

Figura C.8

O p-valor quando $t = -2{,}13$ com 19 graus de liberdade para a alternativa unilateral $\mu < 0$.



Para populações não-normais, o p-valor exato pode ser difícil de ser obtido. No entanto, podemos encontrar p-valores *assimptóticos* usando os mesmos cálculos. Esses p-valores serão válidos para amostras de tamanhos grandes. Para n maior que, digamos, 120, também é possível usar a distribuição normal padrão. A Tabela G.1 é suficientemente detalhada para obtermos p-valores exatos, mas também podemos usar um programa estatístico ou econométrico.

EXEMPLO C.7

(Discriminação Racial na Contratação de Trabalhadores)

Usando os dados de pares comparados do Urban Institute ($n = 241$), obtivemos $t = -4{,}29$. Se Z for uma variável aleatória normal padrão, $P(Z < -4{,}29)$ será, em sentido prático, zero. Em outras palavras, o p-valor (assimptótico) deste exemplo será em essência zero. Isso será evidência bastante forte contra H_0.

SUMÁRIO SOBRE COMO UTILIZAR *p*-VALORES

(i) Escolha um teste estatístico T e decida sobre a natureza da alternativa. Isso determinará se a regra de rejeição será $t > c$, $t < -c$, ou $|t| > c$.

(ii) Use o valor observado da estatística t como o valor crítico e calcule o nível de significância correspondente do teste. Esse será o p-valor. Se a regra de rejeição for da forma $t > c$, então, p-valor $= P(T > t)$. Se a regra de rejeição for $t < -c$, então, p-valor $= P(T < t)$; se a regra de rejeição for $|t| > c$, então, p-valor $= P(|T| > |t|)$.

(iii) Se um nível de significância α tiver sido escolhido, então, rejeitaremos H_0 ao nível $100 \cdot \alpha\%$ se p-valor $< \alpha$. Se p-valor $\geq \alpha$, então, não podemos rejeitar H_0 ao nível $100 \cdot \alpha\%$. Portanto, um p-valor pequeno leva à rejeição de H_0.

A Relação entre Intervalos de Confiança e Testes de Hipóteses

Como tanto a construção de intervalos de confiança como de testes de hipóteses envolvem definições de probabilidade, é natural pensar que eles sejam de alguma forma interligados. Realmente, eles são relacionados. Após um intervalo de confiança ter sido construído, podemos conduzir uma diversidade de testes de hipóteses.

Os intervalos de confiança sobre os quais temos discutido são todos de natureza bilateral. (Neste livro, não teremos a necessidade de construir intervalos de confiança unilaterais.) Assim, intervalos de confiança podem ser usados para testes contra alternativas bilaterais. No caso de uma média populacional, a hipótese nula é dada por (C.31), e a alternativa é (C.34). Suponha que tenhamos construído um intervalo de confiança de 95% para μ. Então, se o valor hipotético de μ sob H_0, μ_0, não for o intervalo de confiança, então, H_0: $\mu = \mu_0$ será rejeitada contra H_1: $\mu \neq \mu_0$ ao nível de 5%. Se μ_0 permanecer nesse intervalo, não poderemos rejeitar H_0 ao nível de 5%. Observe como qualquer valor de μ_0 pode ser testado uma vez tenha sido construído um intervalo de confiança, e como um intervalo de confiança contém mais de um valor, haverá muitas hipóteses nulas que não serão rejeitadas.

EXEMPLO C.8

(Subsídios de Treinamento e Produtividade dos Trabalhadores)

No exemplo de Holzer et al., construímos um intervalo de confiança de 95% para a alteração média na taxa de rejeição μ como $[-2{,}28, -0{,}02]$. Como o zero foi excluído desse intervalo, rejeitamos H_0: $\mu = 0$ contra H_1: $\mu \neq 0$ ao nível de 5%. Esse intervalo de confiança de 95% também significa que não podemos rejeitar H_0: $\mu = -2$ ao nível de 5%. De fato, existirá uma série contínua de hipóteses nulas que não serão rejeitadas, dado esse intervalo de confiança.

Significância Prática *versus* Estatística

Nos exemplos utilizados até agora, produzimos três tipos de evidências concernentes aos parâmetros populacionais: estimativa por ponto, intervalos de confiança e testes de hipóteses. Essas ferramentas para obter informações sobre os parâmetros populacionais são igualmente importantes. Existe uma compreensível tendência dos estudantes de se concentrarem nos intervalos de confiança e nos testes de hipóteses, pois são coisas às quais podemos anexar níveis de confiança ou de significância. Mas em qualquer trabalho precisamos também interpretar as *magnitudes* das estimativas por ponto.

A significância estatística depende do tamanho da estatística t e não apenas do tamanho de $\bar{y}$. Para testar H_0: $\mu = 0$, $t = \bar{y}/\mathrm{ep}(\bar{y})$. Assim, a significância estatística depende da razão de $\bar{y}$ e do seu erro-padrão. Uma estatística t pode ser grande porque $\bar{y}$ é grande ou porque $\mathrm{ep}(\bar{y})$ é pequeno.

EXEMPLO C.9

(Efeito da Largura de Rodovias sobre o Tempo de Viagem)

Seja Y a alteração no tempo de viagem, medida em minutos, de viajantes em uma área metropolitana, do período anterior ao período posterior do alargamento de uma rodovia. Assuma que $Y \sim \text{Normal}(\mu,\sigma^2)$. A hipótese nula de que o alargamento não reduz o tempo de viagem será H_0: $\mu = 0$; a alternativa que ele reduz o tempo de viagem será H_1: $\mu < 0$. Suponha que uma amostra aleatória de viajantes de tamanho $n = 300$ foi obtida para determinar a efetividade do projeto da rodovia. A alteração média no tempo de viagem será computada como $\bar{y} = 3{,}6$, e o desvio-padrão da amostra será $s = 18{,}7$; assim, $\text{ep}(\bar{y}) = 18{,}7/\sqrt{300} \approx 1{,}08$. A estatística t será $-3{,}61/1{,}08 \approx -3{,}33$, que é bem significante estatisticamente; o p-valor será em essência zero. Assim, concluímos que o alargamento da rodovia terá um efeito significante sobre o tempo médio de viagem.

Se o resultado do teste de hipótese for tudo o que foi divulgado sobre o trabalho, ele será enganoso. Divulgar somente a significância estatística mascara o fato de que a redução estimada de 3,6 minutos na média do tempo de viagem é muito pequena. Para sermos honestos, deveremos informar a estimativa por ponto de $-3{,}6$, com o teste de significância.

Embora a magnitude e o sinal da estatística t determinem a significância estatística, a estimativa por ponto $\bar{y}$ determina o que podemos chamar de **significância prática**. Uma estimativa pode ser estatisticamente significante sem ser especialmente grande. Devemos sempre discutir a significância prática em conjunto com a significância estatística da estimativa por ponto; esse tema surgirá com freqüência neste livro.

Encontrar estimativas por ponto que sejam estatisticamente significantes sem ter significância prática freqüentemente ocorre quando trabalhamos com amostras grandes. Para discutir porque isso ocorre, é útil termos a seguinte definição.

CONSISTÊNCIA DE UM TESTE

Um **teste consistente** rejeita H_0 com probabilidade que se aproxime de um conforme o tamanho da amostra cresce, sempre que H_1 for verdadeira.

Uma outra maneira de dizer que um teste é consistente é que, conforme o tamanho da amostra tende ao infinito, o poder do teste se aproxima cada vez mais da unidade, sempre que H_1 for verdadeira. Todos os testes que examinamos neste livro têm essa propriedade. No caso dos testes de hipóteses sobre uma média populacional, a consistência dos testes é uma conseqüência, porque a variância de $\bar{Y}$ converge para zero conforme o tamanho da amostra aumenta. A estatística t para testar H_0: $\mu = 0$ é $T = \bar{Y}/(S/\sqrt{n})$. Como $\text{plim}(\bar{Y}) = \mu$ e $\text{plim}(S) = \sigma$, segue que se, digamos $\mu > 0$, T vai ficando cada vez maior (com alta probabilidade) conforme $n \to \infty$. Em outras palavras, independente de o quanto μ está perto de zero, podemos estar quase certos de que H_0: $\mu = 0$ será rejeitada, devido ao tamanho suficientemente grande da amostra. Isso não fornece nenhuma informação sobre μ ser grande em um sentido prático.

C.7 OBSERVAÇÕES SOBRE NOTAÇÃO

Em nossa revisão de probabilidade e estatística neste capítulo e no Apêndice B, tivemos o cuidado de usar convenções padrão para representar variáveis aleatórias, estimadores e estatísticas de testes. Por exemplo, usamos W para indicar um estimador (variável aleatória) e w para representar uma estimativa específica (resultado da variável aleatória W). Fazer a distinção entre um estimador e uma estimativa é importante para a compreensão de vários conceitos sobre estimação e testes de hipóteses. Porém, fazer essa distinção de forma rápida se tornará um peso na análise econométrica, pois os modelos são mais complicados: muitos parâmetros e variáveis aleatórias estarão envolvidos, e a obediência às convenções habituais da probabilidade e da estatística exigirá muitos símbolos extras.

No texto principal, usamos uma convenção mais simples que é amplamente usada em econometria. Se θ for um parâmetro populacional, a notação $\hat{\theta}$ ("teta chapéu") será usada para representar tanto um estimador como uma estimativa de θ. Essa notação é útil no sentido de que ela propicia uma maneira simples de integrar um estimador ao parâmetro populacional que supostamente ela estará estimando. Assim, se o parâmetro populacional for β, então, $\hat{\beta}$ denotará um estimador ou uma estimativa de β; se o parâmetro for σ^2, $\hat{\sigma}^2$será um estimador ou uma estimativa de σ^2; e assim por diante. Algumas vezes, examinaremos dois estimadores do mesmo parâmetro, caso em que necessitaremos de uma notação diferente como, por exemplo, $\tilde{\theta}$ ("teta til").

Embora o abandono das convenções sobre probabilidade e estatística para indicar estimadores, variáveis aleatórias e estatísticas de testes coloque mais responsabilidade sobre seus ombros, isso não será um grande problema, uma vez que a diferença entre um estimador e uma estimativa seja compreendida. Se estivermos tratando das propriedades *estatísticas* de $\hat{\theta}$ — como, por exemplo, deduzir se ele é ou não não-viesado ou consistente —, então, estaremos necessariamente vendo $\hat{\theta}$ como um estimador. Por outro lado, se escrevermos algo como $\hat{\theta} = 1{,}73$, então, representaremos claramente uma estimativa por ponto a partir de determinada amostra de dados. A confusão que pode surgir com o uso de $\hat{\theta}$ para representar ambas será mínima se você tiver um bom entendimento de probabilidade e de estatística.

RESUMO

Discutimos sobre tópicos de estatística matemática que são muito usados na análise econométrica. A noção de um estimador, que é simplesmente uma regra de combinação de dados para estimar um parâmetro populacional, é fundamental. Tratamos de várias propriedades dos estimadores. As mais importantes propriedades das amostras pequenas são a inexistência de viés e a eficiência, sendo que esta última depende das comparações das variâncias quando os estimadores forem não-viesados. As propriedades das amostras grandes relacionam-se com a seqüência dos estimadores obtidos conforme o tamanho da amostra aumenta, e em econometria depende-se delas. Qualquer estimador de valia é consistente. O teorema do limite central implica que, em amostras grandes, a distribuição amostral da maioria dos estimadores será aproximadamente normal.

A distribuição amostral de um estimador pode ser usada para a construção de intervalos de confiança. Vimos isso na estimação da média de uma distribuição normal e no cálculo de intervalos de confiança aproximados em casos não-normais. O teste de hipótese clássico, que exige a especificação de uma hipótese nula, de uma hipótese alternativa e de um nível de significância, é executado comparando-se uma estatística de teste com um valor crítico. Alternativamente, pode-se calcular um p-valor que nos possibilite conduzir um teste em qualquer nível de significância.

PROBLEMAS

C.1 Sejam Y_1, Y_2, Y_3 e Y_4 variáveis aleatórias independentes e identicamente distribuídas de uma população com média μ e variância σ^2. $\overline{Y} = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$ representa a média dessas quatro variáveis aleatórias.

(i) Quais são o valor esperado e a variância de $\overline{Y}$ em termos de μ e σ^2?

(ii) Agora, considere um estimador diferente de μ:

$$W = \frac{1}{8}Y_1 + \frac{1}{8}Y_2 + \frac{1}{4}Y_3 + \frac{1}{2}Y_4.$$

Esse é um exemplo de uma média ponderada dos Y_i. Mostre que W também é um estimador não-viesado de μ. Encontre a variância de W.

(iii) Com base em suas respostas nas partes (i) e (ii), qual estimador de μ você prefere, $\overline{Y}$ ou W?

C.2 Esta é uma versão mais generalizada do Problema C.1. Sejam Y_1, Y_2, ..., Y_n n variáveis aleatórias não-correlacionadas comparativamente, com média comum μ e variância comum σ^2. Seja $\overline{Y}$ a média da amostra.

(i) Defina a classe dos *estimadores lineares* de μ como

$$W_a = a_1Y_1 + a_2Y_2 + ... + a_nY_n,$$

onde os a_i são constantes. Que restrição sobre os a_i é necessária para que W_a seja um estimador não-viesado de μ?

(ii) Encontre $\text{Var}(W_a)$.

(iii) Para quaisquer números $a_1, a_2, ..., a_n$, a seguinte desigualdade é válida: $(a_1 + a_2 + ... + a_n)^2/n \leq a_1^2 + a_2^2 + ... + a_n^2$. Use isso, com as partes (i) e (ii), para mostrar que $\text{Var}(W_a) \geq \text{Var}(\overline{Y})$ sempre que W_a for não-viesado, de forma que $\overline{Y}$ seja o *melhor estimador linear não-viesado*. [*Sugestão*: Em que se transforma a desigualdade quando a_i satisfaz a restrição da parte (i)?]

C.3 Seja Y a média amostral de uma amostra aleatória com média μ e variância σ^2. Considere dois estimadores alternativos de μ: $W_1 = [(n - 1)/n]\ \overline{Y}$ e $W_2 = \overline{Y}/2$.

(i) Mostre que W_1 e W_2 são ambos estimadores viesados de μ e encontre os vieses. O que acontece com os vieses conforme $n \to \infty$? Comente sobre quaisquer diferenças importantes no viés para os dois estimadores conforme o tamanho da amostra aumenta.

(ii) Encontre os limites de probabilidade de W_1 e W_2. {*Sugestão*: Use as propriedades PLIM.1 e PLIM.2; para W_1, observe que $\text{plim}[(n - 1)/n] = 1$.} Qual estimador é consistente?

(iii) Encontre $\text{Var}(W_1)$ e $\text{Var}(W_2)$.

(iv) Demonstre que W_1 é um estimador melhor que $\overline{Y}$ se μ estiver "próximo" de zero. (Considere tanto o viés como a variância.)

C.4 Para variáveis aleatórias positivas X e Y, suponha que o valor esperado de Y, dado X, seja $E(Y|X) = \theta X$. O parâmetro desconhecido θ mostra como o valor esperado de Y muda com X.

(i) Defina a variável aleatória $Z = Y/X$. Mostre que $E(Z) = \theta$. [*Sugestão*: Use a Propriedade EC.2 em conjunto com a lei das expectativas iteradas, a Propriedade EC.4. Em particular, primeiro mostre que $E(Z|X) = \theta$ e, então, use EC.4.]

(ii) Utilize a parte (i) para provar que o estimador $W_1 = n^{-1}\sum_{i=1}^{n}(Y_i/X_i)$ é não-viesado para θ, quando $\{(X_i,Y_i): i = 1, 2, ..., n\}$ for uma amostra aleatória.

(iii) Explique por que o estimador $W_2 = \overline{Y}/\overline{X}$, onde as barras superiores significam médias das amostras, não é o mesmo que W_1. Apesar disso, mostre que W_2 também é não-viesado para θ.

(iv) A tabela seguinte contém dados sobre a produção de milho de diversas regiões de Iowa. O United States Department of Agriculture (USDA) faz a previsão dos hectares[1] de milho em cada região com base em fotos de satélite. Os pesquisadores contam o número de "pixels[2]" de milho na foto do satélite (em oposição a, por exemplo, número de pixels de soja ou de terra não cultivada) e usam esses números para prognosticar o número efetivo de hectares. Para desenvolver uma equação de previsão para ser usada de forma generalizada para as regiões, o USDA entrevistou agricultores em regiões selecionadas para obter a produção de milho em hectares. Seja Y_i = produção de milho na região i e X_i = número de pixels de milho na foto do satélite da região i. Existem $n = 17$ observações de oito regiões. Utilize essa amostra para computar as estimativas de θ desenvolvidas nas partes (ii) e (iii). As estimativas são semelhantes?

Imagem	Produção de Milho	Pixels de Milho
1	165,76	374
2	96,32	209
3	76,08	253
4	185,35	432
5	116,43	367
6	162,08	361
7	152,04	288
8	161,75	369
9	92,88	206
10	149,94	316
11	64,75	145
12	127,07	355
13	133,55	295
14	77,70	223
15	206,39	459
16	108,33	290
17	118,17	307

[1] Medida agrária igual a 10.000 metros quadrados. (N. do T.)

[2] Unidade de informação que descreve um ponto em uma imagem gráfica computadorizada; é o menor ponto de luz cuja cor e luminosidade podem ser controlados na tela. (N. do T.)

C.5 Seja Y uma variável aleatória de Bernoulli(θ) com $0 < \theta < 1$. Suponha que estamos interessados em estimar a razão de probabilidades, $\gamma = \theta/(1 - \theta)$, que é a probabilidade de êxito sobre a probabilidade de fracasso. Dada uma amostra aleatória $\{Y_1, ..., Y_n\}$, sabemos que um estimador consistente e não-viesado de θ é $\overline{Y}$, a proporção de êxitos em n tentativas. Um estimador natural de γ é $G = \overline{Y}/(1 - \overline{Y})$, a proporção de êxitos sobre a proporção de fracassos na amostra.

(i) Por que G não é um estimador não-viesado de γ?

(ii) Use PLIM.2(iii) para mostrar que G é um estimador consistente de γ.

C.6 Você foi contratado pelo governador para examinar se um imposto sobre bebidas alcoólicas reduziu o consumo de bebidas alcoólicas em seu estado. Você tem condições de obter, para uma amostra de pessoas selecionadas aleatoriamente, a diferença no consumo de bebidas alcoólicas (em onças) dos anos anterior e posterior à instituição do imposto. Para $i^{\text{-ésima}}$ pessoa que foi extraída aleatoriamente da amostra da população, Y_i representa a alteração no consumo de bebidas alcoólicas. Trate-as como uma amostra aleatória de uma distribuição Normal(μ,σ^2).

(i) A hipótese nula é que não houve mudança na média de consumo de bebidas alcoólicas. Represente isso formalmente, em termos de μ.

(ii) A hipótese alternativa é que houve um declínio no consumo de bebidas alcoólicas; estabeleça a hipótese alternativa em termos de μ.

(iii) Agora suponha que o tamanho de sua amostra seja $n = 900$ e que você obtenha as estimativas $\overline{y} = -32{,}8$ e $s = 466{,}4$. Calcule a estatística t do teste de H_0 contra H_1; obtenha o p-valor do teste. (Devido ao grande tamanho da amostra, use apenas a distribuição normal padrão tabulada na Tabela G.1.) Você rejeita H_0 ao nível de 5%? E ao nível de 1%?

(iv) Você diria que a queda estimada do consumo é grande, em magnitude? Comente sobre a significância prática *versus* a significância estatística dessa estimativa.

(v) O que foi implicitamente assumido em sua análise sobre outros determinantes do consumo de bebidas alcoólicas ao longo do período de dois anos, para inferir causalidade entre a alteração do imposto e o consumo de bebidas alcoólicas?

C.7 A nova administração de uma padaria alega que os trabalhadores agora são mais produtivos do que eram sob a administração anterior, razão pela qual os salários foram "aumentados de forma geral". Sejam W_i^b o salário do trabalhador i sob a administração antiga e o salário do trabalhador i após a mudança. A diferença será $D_i \equiv W_i^a - W_i^b$. Assuma que os D_i são uma amostra aleatória de uma distribuição Normal(μ,σ^2).

(i) Usando os dados seguintes de 15 trabalhadores, construa um intervalo de confiança exato de 95% para μ.

(ii) Escreva formalmente a hipótese nula de que não houve alteração na média dos salários. Em particular, qual é o E(D_i) de acordo com H_0? Se você fosse contratado para verificar a validade da alegação da nova administração, qual seria a hipótese alternativa relevante em termos de $\mu = \text{E}(D_i)$?

(iii) Teste a hipótese nula da parte (ii) contra a alternativa declarada, aos níveis de 5% e 1%.

(iv) Obtenha o p-valor do teste na parte (iii).

Trabalhador	Salário Antes	Salário Depois
1	8,30	9,25
2	9,40	9,00

(Continua...)

(...continuação)

Trabalhador	Salário Antes	Salário Depois
3	9,00	9,25
4	10,50	10,00
5	11,40	12,00
6	8,75	9,50
7	10,00	10,25
8	9,50	9,50
9	10,80	11,50
10	12,55	13,10
11	12,00	11,50
12	8,65	9,00
13	7,75	7,75
14	11,25	11,50
15	12,65	13,00

C.8 O jornal *The New York Times* (05/02/1990) publicou a atuação dos dez melhores arremessadores de três pontos da NBA. A tabela seguinte resume esses dados:

Jogador	AT – AC
Mark Price	429-188
Trent Tucker	833-345
Dale Ellis	1.149-472
Craig Hodges	1.016-396
Danny Ainge	1.051-406
Byron Scott	676-260
Reggie Miller	416-159
Larry Bird	1.206-455
Jon Sundvold	440-166
Brian Tayllor	417-157

Nota: AT = arremessos tentados e AC = arremessos convertidos.

Para um determinado jogador, o resultado de um arremesso específico pode ser modelado como uma variável de Bernoulli (zero-um): se Y_i for o resultado do arremesso i, então, $Y_i = 1$ se o arremesso for convertido, e $Y_i = 0$ se o arremesso foi perdido. Seja θ a probabilidade de conversão de qualquer determinada tentativa de arremesso de três pontos. O estimador natural de θ será $\overline{Y} = AC/AT$.

(i) Estime θ para Mark Price.

(ii) Encontre o desvio-padrão do estimador $\overline{Y}$ em termos de θ e o número de arremessos tentados, n.

(iii) A distribuição assimptótica de $(Y - \theta)/\text{ep}(\overline{Y})$ será normal padrão quando $\text{ep}(\overline{Y}) = \sqrt{\overline{Y}(1 - \overline{Y})/n}$. Use esse fato para testar H_0: $\theta = 0{,}5$ contra H_1: $\theta < 0{,}5$ para Mark Price. Use um nível de significância de 1%.

C.9 Suponha que um ditador militar de um país sem nome promova um plebiscito (um voto de confiança sim/não) e afirme que teve o apoio de 65% dos votantes. Um grupo de direitos humanos suspeita que houve jogo sujo e contrata você para verificar a validade da afirmação do ditador. Você tem um orçamento que lhe possibilita fazer uma amostragem aleatória de 200 votantes no país.

(i) Defina X como o número de votos sim obtidos da amostra aleatória de 200 de toda a população votante. Qual será o valor esperado de X se, realmente, 65% de todos os votantes apoiaram o ditador?

(ii) Qual será o desvio-padrão de X, novamente assumindo que a fração verdadeira de votos sim tenha sido 0,65?

(iii) Agora, você coleta sua amostra de 200, e descobre que 115 pessoas efetivamente votaram sim. Use o TLC para aproximar a probabilidade de que você encontraria 115 ou menos votos sim de uma amostra aleatória de 200 se, realmente, 65% de toda a população tivesse votado sim.

(iv) Como você explicaria a relevância do número na parte (iii) para alguém que não tem conhecimento de estatística?

C.10 Antes de uma greve ter prematuramente terminado com a temporada de 1994 da liga principal de beisebol, Tony Gwynn, do San Diego Padres, tinha 165 rebatidas válidas em 419 rebatidas, para uma média de rebatidas de 0,419. Houve muita discussão se Gwynn seria um rebatedor potencial de 0,400 rebatidas válidas naquele ano. Esse problema pode ser expresso em termos da probabilidade de Gwynn fazer uma rebatida válida em uma de suas oportunidades de rebater; vamos chamá-la θ. Seja Y_i o indicador Bernoulli(θ) igual a unidade se Gwynn fizer uma rebatida válida na sua i-ésima vez de rebater, e zero, caso contrário. Então, $Y_1, Y_2, ..., Y_n$ será uma amostra aleatória de uma distribuição de Bernoulli, onde θ será a probabilidade de êxito, e $n = 419$.

Nossa melhor estimativa por ponto de θ será a média de rebatidas de Gwynn, que simplesmente será a proporção dos êxitos: $\overline{y} = 0{,}394$. Usando o fato de que $\text{ep}(\overline{y}) = \sqrt{\overline{y}(1 - \overline{y})/n}$, construa um intervalo de confiança aproximado de 95% para θ, usando a distribuição normal padrão. Você diria que há forte evidência contra a possibilidade de que Gwynn teria sido um rebatedor potencial de 0,400 rebatidas válidas? Explique.