4.8.1. Estimação de proporções (amostras grandes)

4.8.1.1. Distribuição amostral da proporção

Na seção **4.4** apresentamos o conceito de *distribuição amostral*, aplicado à proporção de sucessos encontrados nas amostras tiradas de uma população. Esta proporção é uma *variável aleatória*: como as amostras são obtidas por meio de sorteios (pelo menos em princípio), cada amostra tem uma proporção de sucessos diferente e podemos usar modelos probabilísticos para estudar como variam estas proporções. Na seção **4.4.1** vimos um teorema afirmando que, para amostras grandes, a *distribuição amostral da proporção* tende para um modelo normal. Reproduzimos abaixo este teorema.

Teorema 1: *Distribuição amostral da proporção de sucessos P*
Se retiramos amostras aleatórias grandes de uma população infinita com proporção de sucessos π, a proporção de sucessos P=X/n nas amostras terá distribuição que tende para a normal:

$$P \rightarrow N(\mu_P, \sigma_P^2) \text{ quando } n \rightarrow \infty$$

onde: $\mu_P = \pi$ e $\sigma_P^2 = \frac{\pi(1-\pi)}{n} \rightarrow \sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$

Portanto, se conhecemos o parâmetro π de uma população (isto é, sua proporção de sucessos), podemos por meio do modelo normal prever em quais intervalos é mais provável encontrarmos as proporções *P* de sucessos encontradas em amostras aleatórias extraídas desta população.

Voltemos, por exemplo, ao item (iii) da seção **4.4.4**, no qual analisamos uma amostra de 495 crianças nascidas em um hospital de São Paulo. Se na população a proporção de nascimentos de *meninos* for igual à de *meninas*, isto é:

P(meninos) = P(meninas) = π = 0,5

podemos deduzir, usando o modelo normal, um intervalo que tem 0,95 de probabilidade de conter a proporção P de sucessos numa amostra aleatória de 495 crianças.

Para obtermos uma probabilidade de 0,95 na parte central da curva, limitamos de cada lado da média uma área de 0,95 /2 = 0,475 (Fig. **1**):

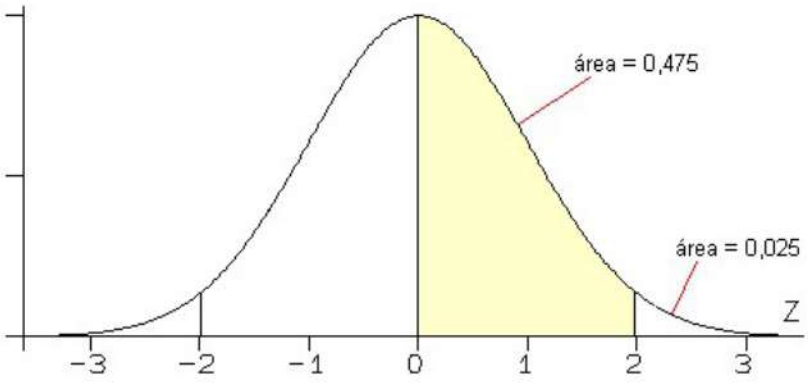

Figura 1. Curva normal com área simétrica de 0,95 delimitada na sua parte central

Na tabela da curva normal, vemos que o limite superior da área de 0,475 à direita da média é o valor de Z igual a:

$$z = +1{,}96$$

Como a distribuição é simétrica, o valor de Z correspondente ao limite do lado esquerdo será o mesmo, com o sinal trocado:

$$z = -1{,}96$$

Segundo o Teorema **1**, a distribuição normal terá média e desvio-padrão dados por

$$\mu_P = \pi = 0{,}5$$

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0{,}5(1-0{,}5)}{495}} = 0{,}0225$$

O limite superior do intervalo desejado é portanto:

$$\begin{aligned} P_{sup} &= \pi + 1{,}96\sigma_P \\ &= 0{,}5 + 1{,}96 \times 0{,}0225 = 0{,}5441 \end{aligned}$$

O limite inferior é:

$$\begin{aligned} P_{inf} &= \pi - 1{,}96\sigma_P \\ &= 0{,}5 - 1{,}96 \times 0{,}0225 = 0{,}4559 \end{aligned}$$

Na amostra, foram encontrados 260 meninos, o que dá uma proporção de

$$P = 260 / 495 = 0{,}5252$$

Como a proporção de meninos encontrada na amostra ficou dentro do intervalo que calculamos supondo a hipótese de que a P(meninos)=P(meninas)=0,5, concluímos que não há nenhuma evidência que nos leve a rejeitar a hipótese.

Generalizando, podemos dizer que se amostras aleatórias grandes forem retiradas de uma população cuja proporção de sucessos é π, há uma probabilidade de 0,95 de as proporções P encontradas nas amostras estarem dentro do intervalo limitado pelos valores:

$$P_{inf} = \pi - 1{,}96\sigma_P$$
$$P_{sup} = \pi + 1{,}96\sigma_P$$

Ou seja,

$$P(\pi - 1{,}96\sigma_P \le \pi \le \pi + 1{,}96\sigma_P) = 0{,}95 \quad (1)$$

onde $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$ (2)

Podemos escrever este intervalo mais simplesmente como:

$$\pi \pm 1{,}96\sigma_P \rightarrow \pi \pm 1{,}96\sqrt{\frac{\pi(1-\pi)}{n}} \quad (3)$$

Vejamos agora o caminho contrário. Suponha que você queira estimar a proporção de sucessos π numa população, mas não tenha nenhuma hipótese inicial para testar. Por exemplo, você quer estimar a proporção π de eleitores que apóiam o seu candidato, antes de uma eleição. Se você tira uma amostra e encontra uma proporção P de eleitores favoráveis, como pode, a partir deste P, fazer uma estimativa de π? Esta estimativa será feita por meio de um *intervalo de confiança*, a técnica de *Inferência Estatística* que veremos a seguir.

4.8.1.2. Estimação por meio de intervalo de confiança

No item anterior, delimitamos um intervalo ao redor da proporção populacional π que tem probabilidade conhecida de conter a proporção P de uma amostra. Faremos agora caminho contrário: dada a proporção P que encontramos numa amostra, calcularemos em torno dela um intervalo que tem uma probabilidade conhecida de conter a proporção π da população.

O Teorema **1** afirma que a proporção P de sucessos nas amostras tem distribuição amostral gaussiana, e podemos portanto afirmar que há 0,95 de probabilidade deste P se encontrar dentro do intervalo definido pelas equações (**1**) e (**2**), deduzidas na seção anterior. Substituindo σ_P na eq. (**1**) por seu valor calculado na eq. (**2**), obtemos:

$$P\left[\pi - 1{,}96\sqrt{\frac{\pi(1-\pi)}{n}} < P < \pi + 1{,}96\sqrt{\frac{\pi(1-\pi)}{n}}\right] = 0{,}95$$

Considerando o lado esquerdo desta desigualdade:

$$\pi - 1{,}96\sqrt{\frac{\pi(1-\pi)}{n}} < P$$
$$\mu < P + 1{,}96\sqrt{\frac{\pi(1-\pi)}{n}} \quad (4)$$

Considerando agora o lado direito:

$$P < \pi + 1{,}96\sqrt{\frac{\pi(1-\pi)}{n}}$$

$$P - 1{,}96\sqrt{\frac{\pi(1-\pi)}{n}} < \pi \qquad (5)$$

Reunindo as duas desigualdades em (4) e (5):

$$P - 1{,}96\sqrt{\frac{\pi(1-\pi)}{n}} < \mu < P + 1{,}96\sqrt{\frac{\pi(1-\pi)}{n}} \qquad (6)$$

Note que o valor de π é obviamente desconhecido (é o que estamos tentando descobrir!). Para construir o intervalo, usaremos P como uma estimativa de π (uma estimativa *pontual*; veja seção **4.8**). Teremos então o desvio-padrão dado por:

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} \approx \sqrt{\frac{P(1-P)}{n}} \qquad (7)$$

Inserindo em (5) o desvio-padrão estimado em (6), chegamos ao intervalo:

$$P - 1{,}96\sqrt{\frac{P(1-P)}{n}} < \mu < P + 1{,}96\sqrt{\frac{P(1-P)}{n}} \qquad (8)$$

Este intervalo resultante é chamado de *Intervalo de Confiança* de 0,95 para a proporção populacional π. A notação deste intervalo não é padronizada, e varia um pouco de livro para livro; iremos usar a notação:

$$IC_{\pi}^{0,95}: \; P \pm 1{,}96\sqrt{\frac{P(1-P)}{n}} \qquad (9)$$

Note que as expressões em (3) e (9), reproduzidas abaixo, são muito parecidas:

$$\pi \pm 1{,}96\sqrt{\frac{\pi(1-\pi)}{n}}$$

$$P \pm 1{,}96\sqrt{\frac{P(1-P)}{n}}$$

A eq. (3) dá um intervalo em torno de π que tem 0,95 de probabilidade de conter P; a eq. (9) dá um intervalo em torno de P, que tem 0,95 de probabilidade de conter π.

Isto é fácil de entender intuitivamente. Suponha que duas cidades A e B estejam a menos de 10 km de distância uma da outra. Se você sabe onde está a cidade A, pode fazer em torno dela um círculo de raio igual a 10 km, e achará B dentro deste círculo; se sabe onde está a cidade B, pode fazer um círculo em torno dela, e achará A dentro deste círculo. (A diferença é que, no caso da Estatística, não há *certeza*; o que podemos dizer é que os valores de P e π *provavelmente* estarão perto um do outro).

4.8.1.3. Observações sobre a estimação por meio de intervalos de confiança

Há algumas observações importantes que devem ser feitas sobre a estimação por meio de intervalos de confiança.

(i) Nível de confiança do IC

O IC calculado pela expressão (**9**) tem uma probabilidade de 0,95 de conter o valor verdadeiro de π; esta probabilidade é chamada de *nível de confiança* do intervalo. Se queremos usar um nível de confiança diferente, temos que trocar o valor 1,96 pelo valor de Z correspondente à probabilidade desejada. Podemos por exemplo usar um intervalo que tem nível de confiança de 0,80; na tabela da curva normal, vemos que o Z correspondente é 1,28 e portanto o IC será dado por:

$$IC_{\pi}^{0,80}: \; P \pm 1{,}28\sqrt{\frac{P(1-P)}{n}} \qquad (10)$$

(ii) Margem de erro e precisão de uma estimativa

Nas expressões (**9**) e (**10**), os termos

$$1{,}96\sqrt{\frac{P(1-P)}{n}} \qquad \text{e} \qquad 1{,}28\sqrt{\frac{P(1-P)}{n}}$$

dão as *margens de erro* destas duas estimativas. Quanto menor este margem, mais *precisa* é a estimativa. Por exemplo, se há duas estimativas para a proporção de eleitores que apoiam um candidato numa eleição:

$$(0{,}35 \pm 0{,}01) \qquad (0{,}35 \pm 0{,}03)$$

a primeira é mais precisa do que a segunda, porque sua margem de erro é menor.

Em geral, queremos que uma estimativa seja o mais precisa possível. A estimativa obtida em (**10**) é mais precisa do que a obtida em (**9**); porém, o preço que pagamos pelo aumento da precisão foi a perda de confiança: o nível de confiança agora é de apenas 0,80; há 0,20 de probabilidade de a estimativa em (**10**) estar errada, contra apenas 0,05 da estimativa em (**9**).

(iii) Determinação do tamanho da amostra

O ideal é aumentarmos a precisão da estimativa sem diminuir a confiança do intervalo. Na eq. (**9**), vemos que a margem de erro, além de depender do valor "1,96", depende também dos valores de P e de n. O valor de P, evidentemente, não está sob nosso controle; o tamanho n da amostra porém está – quanto maior a amostra que usarmos, menor a margem de erro.

Suponhamos o exemplo mencionado no item anterior, sobre a estimativa da proporção de eleitores que apoiam um candidato numa eleição. Para obtermos uma estimativa (0,35 ± 0,03), com confiança de 0,95, precisamos fazer com que a margem de erro em (**9**) seja igual a 0,03:

$$1{,}96\sqrt{\frac{0{,}35(1-0{,}35)}{n}} = 0{,}03 \quad \rightarrow \quad n \cong 971$$

Se quisermos uma estimativa mais precisa, com margem de erro de ±0,01, o tamanho da amostra será aumentado para n ≅ 8739. A precisão não é uma função linear do tamanho da amostra; para triplicar a precisão, temos que usar uma amostra 9 vezes maior.

Note que fizemos o cálculo acima supondo que P = 0,35. Se não temos nenhuma estimativa, mesmo aproximada, de P (por exemplo, alguma estimativa obtida em pesquisa feita anteriormente), podemos usar P = 0,5. Este é o pior caso, pois então as amostras terão que ser maiores: para as duas margens de erro consideradas acima, os tamanhos serão de n=1067 e n=9604, respectivamente.

(iv) Interpretação do IC

O IC calculado pela expressão (**9**) é um intervalo cujo nível de confiança é igual a 0,95. Isto quer dizer que, se calcularmos um intervalo destes, ele terá 0,95 de probabilidade de conter o valor real de π. Se tirarmos um grande número de amostras de uma população e calcularmos ICs a partir delas, em média 95% destes ICs (19 em cada 20 ICs) conterão o valor real de π; 5% deles, porém (1 em cada 20 ICs) não conterão o valor real.

O mesmo acontece com os *testes de hipótese* que vimos em seções anteriores. Se tiramos 100 amostras de uma população e fazemos testes de um parâmetro usando α=0,05, em média chegaremos à conclusões erradas em 5 destas amostras. Em qualquer técnica de Inferência Estatística, sempre existe a probabilidade de chegarmos a um resultado errado; podemos reduzir esta probabilidade, mas nunca eliminá-la.

(v) É preciso levar em conta o tipo de amostra usado

É preciso que tenhamos sempre em mente um ponto muito importante: todas estas fórmulas derivadas do Teorema **1** *só servem se a amostra for aleatória simples!* Uma amostra aleatória simples é aquela em que todos os eleitores são numerados, e depois alguns deles são sorteados para fazerem parte da amostra. Na prática, isto nem sempre é feito desta maneira; frequentemente são usados outros tipos de amostra (amostras *estratificadas*, por exemplo), para os quais as distribuições amostrais seguem modelos diferentes. Estes tipos de amostra e as distribuições amostrais de suas estatísticas são estudados na área de *Amostragem*, que em geral não é abordada em cursos introdutórios de Estatística.

4.8.1.4. Exemplos

(i) Razão de sexo : probabilidade de nascimento de meninos (1)

Voltemos ao problema estudado na seção **4.4.3.1**. Numa amostra de 495 registrados em São Paulo, foram encontrados 260 meninos. Partindo desta amostra, que estimativa podemos fazer da probabilidade π de nascimento de meninos?

Os dados são:

n = 495 (tamanho da amostra)
X = 260 (número de meninos na amostra)

A proporção de meninos na amostra é portanto:

$$P = 260 / 495 = 0{,}5252$$

Podemos usar a expressão em (**9**) para calcular um intervalo de confiança para a probabilidade π:

$$IC_{\pi}^{0,95}: \; P \pm 1{,}96\sqrt{\frac{P(1-P)}{n}}$$

$$IC_{\pi}^{0,95}: \; 0{,}5252 \pm 1{,}96\sqrt{\frac{0{,}5252(1-0{,}5252)}{495}}$$

$$IC_{\pi}^{0,95}: \; 0{,}48 \; a \; 0{,}57$$

Note que não podemos, a partir dos dados desta amostra, afirmar que a probabilidade de nascimento de meninos seja *maior* do que a das meninas; não podemos concluir se é maior, ou menor, ou igual. No teste de hipóteses que fizemos na seção **4.4.3.1** concluímos que o resultado obtido nesta amostra era *não significativo*, e o que encontramos neste IC corrobora esta conclusão.

(ii) Razão de sexo : probabilidade de nascimento de meninos (2)

Voltemos ao problema estudado na seção **4.4.3.2**. Entre 4.065.014 nascimentos nos EUA em 1992, houve 2.081.287 meninos. Partindo desta amostra, que estimativa podemos fazer da probabilidade π de nascimento de meninos?

Os dados são:

n =	4065014	(tamanho da amostra)
X =	2081287	(número de meninos na amostra)

A proporção de meninos na amostra é portanto:

P = 2081287 / 4065014 = 0,512

Usando a expressão em (**9**) para calcular um intervalo de confiança para a probabilidade π:

$$IC_{\pi}^{0,95}: \; P \pm 1{,}96\sqrt{\frac{P(1-P)}{n}}$$

$$IC_{\pi}^{0,95}: \; 0{,}512 \pm 1{,}96\sqrt{\frac{0{,}512(1-0{,}512)}{4065014}}$$

$$IC_{\pi}^{0,95}: \; 0{,}5115 \; a \; 0{,}5125$$

Agora podemos, a partir dos dados desta amostra, afirmar (com 0,95 de confiança) que a probabilidade de nascimento de meninos é realmente *maior* do que a das meninas. No teste de hipóteses que fizemos na seção **4.4.3.2** concluímos que o resultado obtido nesta amostra era *significativo* (valor-p ≅ 0,0000); o IC portanto corrobora esta conclusão. Note que esta estimativa é muito mais precisa do que a do item anterior: sua margem de erro agora é de apenas ± 0.00097 (contra a anterior de ±0.088), porque o tamanho da amostra empregada é muitas vezes maior.

(iii) Lançamentos de tachinhas.

Por último, voltemos ao problema estudado na seção **4.4.3.3**. Em 1000 lançamentos, as tachinhas cairam com a ponta para cima 548 vezes. Iremos estimar por um IC a probabilidade de uma tachinha cair com a ponta para cima:

Os dados são:

n =	1000	(tamanho da amostra)
X =	548	(número de tachinhas caídas com ponta para cima)

A proporção de tachinhas com a ponta para cima foi na amostra é:

P = 548 / 1000 = 0,548

Usando a expressão em (**9**) para calcular um intervalo de confiança:

$$IC_{\pi}^{0,95}: \; P \pm 1{,}96\sqrt{\frac{P(1-P)}{n}}$$

$$IC_{\pi}^{0,95}: \; 0{,}548 \pm 1{,}96\sqrt{\frac{0{,}548(1-0{,}548)}{1000}}$$

$$IC_{\pi}^{0,95}: \; 0{,}517 \; a \; 0{,}579$$

Na seção **4.4.3.3** concluímos que o resultado obtido nesta amostra era *significativo* (valor-p ≅ 0,0024), e o IC corrobora esta conclusão; há evidência de que probabilidade de a tachinha cair com a ponta para cima é maior do que a de cair com a ponta para baixo.