CPEQS

AULA 03
Estimativas e tamanhos amostrais

Ernesto F. L. Amaral

03 de outubro de 2013

Centro de Pesquisas Quantitativas em Ciências Sociais (CPEQS)
Faculdade de Filosofia e Ciências Humanas (FAFICH)
Universidade Federal de Minas Gerais (UFMG)

Fonte:

Triola, Mario F. 2008. "Introdução à estatística". 10 ª ed. Rio de Janeiro: LTC. Capítulo 7 (pp.250-303).

ESQUEMA DA AULA

– Estimação da proporção populacional.

– Estimação da média populacional: σ conhecido.

– Estimação da média populacional: σ desconhecido.

OBJETIVO DO CAPÍTULO

– Neste capítulo, são usados dados de amostras aleatórias simples para obter estimativas de parâmetros populacionais, o que é a essência da inferência estatística.

– As duas principais aplicações da inferência estatística envolvem o uso de dados amostrais para:

  – **Estimar o valor de um parâmetro populacional** (proporções e médias).

  – Testar alguma afirmação (ou hipótese) sobre uma população.

– São ainda apresentados métodos para determinação das margens de erro, dos tamanhos amostrais e dos **intervalos de confiança** necessários para estimar esses parâmetros.

ESTIMAÇÃO DA PROPORÇÃO POPULACIONAL

ESTIMAÇÃO DA PROPORÇÃO POPULACIONAL

– A intenção é de usar uma proporção amostral para estimar o valor de uma proporção populacional com um intervalo de confiança.

– São apresentados métodos para encontrar o tamanho da amostra necessário para estimar a proporção populacional.

– É importante:

  – Entender o que são, o que fazem e por que são necessários os intervalos de confiança.

  – Desenvolver a habilidade de construir estimativas de intervalos de confiança de proporções populacionais.

  – Aprender como interpretar corretamente um intervalo de confiança.

NOTAÇÃO PARA PROPORÇÕES

– *p* = proporção populacional.

– $\hat{p} = x/n$ = proporção amostral de *x* sucessos em uma amostra de tamanho *n*.

– $\hat{q} = 1 - \hat{p}$ = proporção amostral de fracassos em uma amostra de tamanho *n*.

– Esta seção se concentra na proporção populacional *p*, que é o mesmo que trabalhar com probabilidades e porcentagens.

– Expresse porcentagens em forma decimal.

ESTIMATIVA PONTUAL

– Se desejamos estimar proporção populacional com único valor, a melhor estimativa é $\hat{p}$ (estimativa pontual).

– Estimativa pontual é um único valor usado para aproximar um parâmetro populacional.

– Proporção amostral $\hat{p}$ é a melhor estimativa pontual da proporção populacional *p*.

– A estimativa pontual é usada porque é não-viesado e é o mais consistente dos estimadores que poderiam ser usados:

  – Distribuição das proporções amostrais tende a centralizar em torno do valor de *p*.

  – Proporções amostrais não subestimam/superestimam *p*.

  – Desvio padrão das proporções amostrais tende a ser menor do que desvios padrões de outros estimadores.

POR QUE USAR INTERVALOS DE CONFIANÇA?

– Como a estimativa pontual não diz o quão precisa ela é, os estatísticos desenvolveram o intervalo de confiança (estimativa intervalar).

– **Intervalo de confiança (IC)** é uma faixa (ou intervalo) de valores usada para estimar o verdadeiro valor de um parâmetro populacional.

– A um intervalo de confiança é associado um nível de confiança, por exemplo, 0,95 (ou 95%).

– O **nível de confiança (NC)** apresenta a taxa de sucesso do procedimento usado para construir o intervalo de confiança.

– Nível de confiança é expresso como probabilidade ou área ($1-\alpha$), em que **α é o nível de significância**.

– Quanto maior o NC, maior o IC.

NÍVEL DE CONFIANÇA

– **Nível de confiança** (grau de confiança ou coeficiente de confiança) é a proporção de vezes que o intervalo de confiança realmente contém o parâmetro populacional, supondo que processo seja repetido várias vezes.

– As escolhas mais comuns para nível de confiança são 90% (α = 0,10), 95% (α = 0,05) e 99% (α = 0,01).

– Escolha de 95% é mais comum porque resulta em bom equilíbrio entre **precisão** (largura do intervalo de confiança) e **confiabilidade** (nível de confiança).

– Precisão (exatidão) é a qualidade de que o resultado da amostra reflita o mundo real.

– Confiabilidade é a qualidade de uma determinada técnica produzir os mesmos resultados em várias aplicações.

INTERPRETAÇÃO DE INTERVALO DE CONFIANÇA

– Por exemplo: $n = 280$; $0,381 < p < 0,497$.

– **Correto:** estamos 95% confiantes de que o intervalo de 0,381 a 0,497 realmente contém o verdadeiro valor de *p*.

  – Se selecionássemos muitas diferentes amostras de tamanho 280 e construíssemos os intervalos de confiança correspondentes, 95% deles realmente conteriam o valor da proporção populacional *p*.

  – O nível de 95% se refere à taxa de sucesso do processo em uso para se estimar a proporção populacional, e não se refere à própria proporção populacional.

– **Errado:** como o valor de *p* é fixo, é incorreto dizer que há uma chance de 95% de que o verdadeiro valor de *p* esteja entre 0,381 e 0,497.

VALORES CRÍTICOS

– O escore padrão *z* ou valor crítico ($z_{\alpha/2}$) separa proporções amostrais que têm chance de ocorrer das que não têm.

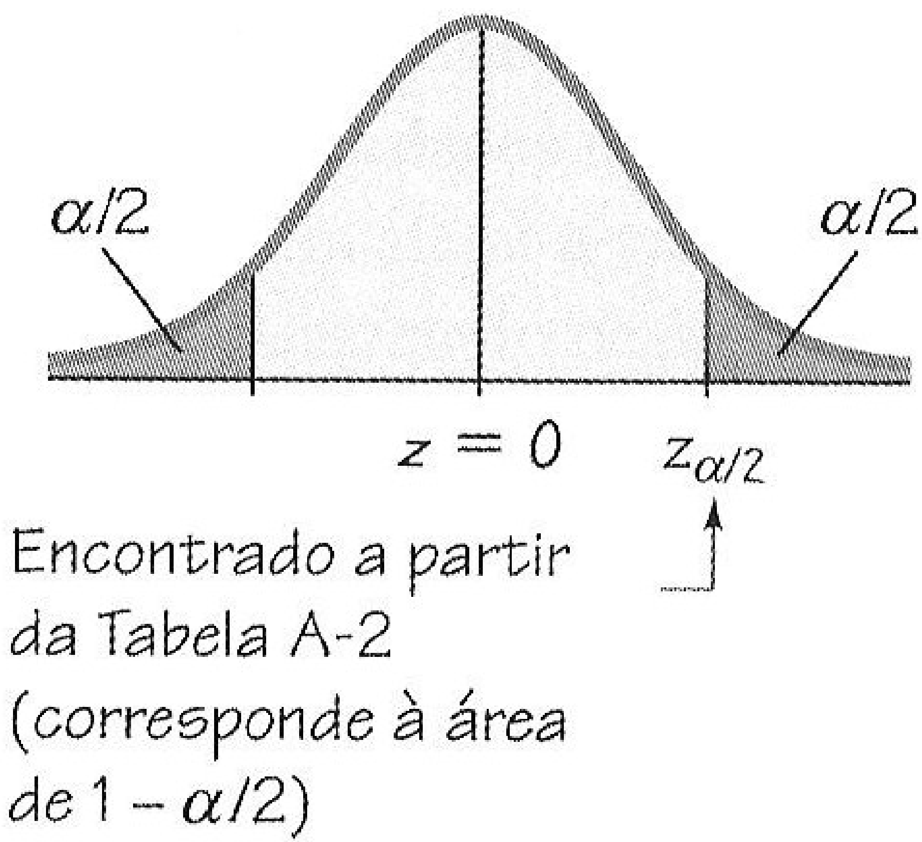
– Os valores críticos se baseiam nestas observações:

- – A distribuição amostral das proporções amostrais pode ser aproximada por uma distribuição normal.
- – Proporções amostrais têm uma chance relativamente pequena de cair em uma das caudas da curva normal.
- – Representando cada cauda por $\alpha/2$, há uma probabilidade total α de que uma proporção amostral caia em uma das duas caudas.
- – Há uma probabilidade de $1-\alpha$ de que uma proporção amostral caia na região entre os pontos críticos (+ e –).

VALORES CRÍTICOS NA DISTRIBUIÇÃO NORMAL PADRÃO

– Valor crítico é um número que separa estatísticas amostrais que têm chance de ocorrer daquelas que não têm.

– O número $z_{\alpha/2}$ é um valor crítico que separa uma área $\alpha/2$ na cauda direita da distribuição normal padronizada.

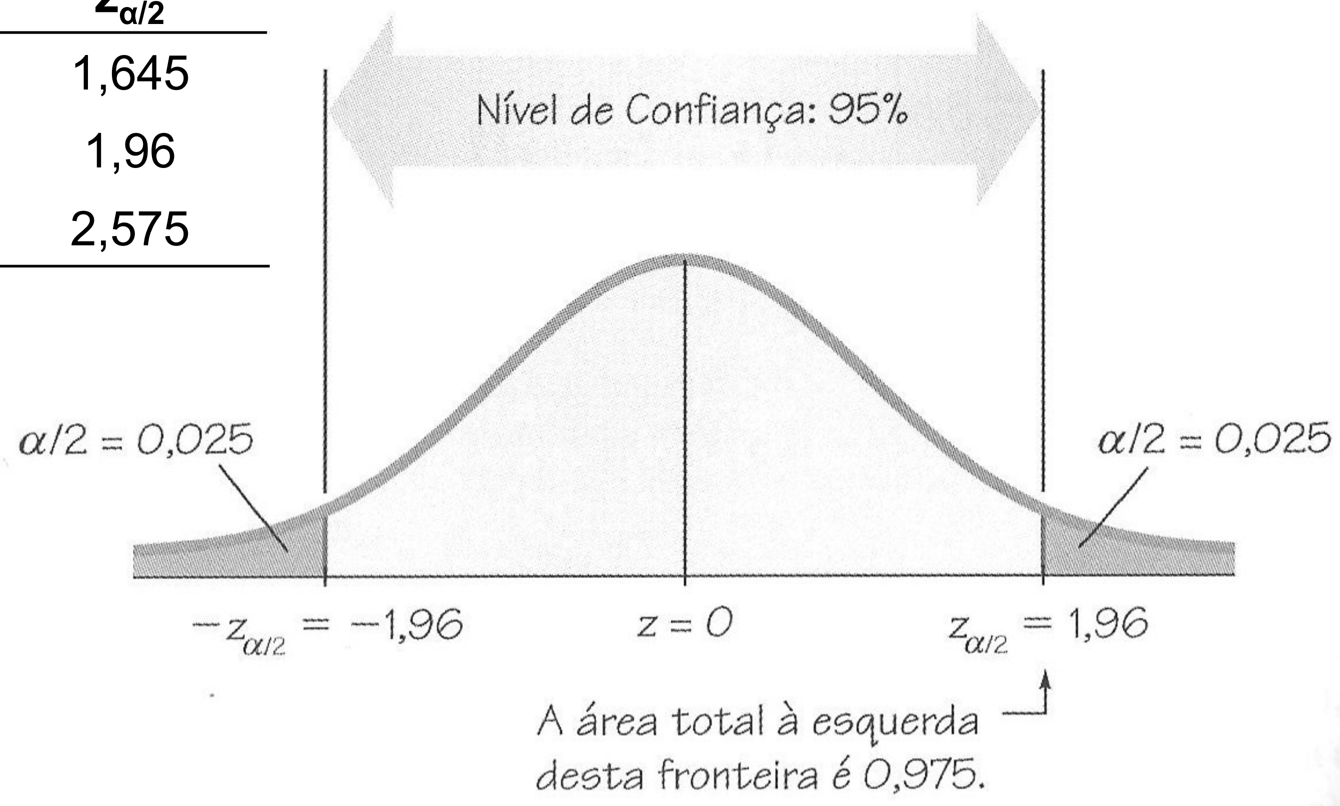


MAIS SOBRE VALORES CRÍTICOS

– O valor de $z_{\alpha/2}$ está na fronteira da cauda direita e o valor de $-z_{\alpha/2}$ está na fronteira da cauda da esquerda.

– Encontrando $z_{\alpha/2}$ para um nível de confiança específico...

Nível de confiança	**α**	**Valor crítico $z_{\alpha/2}$**
90%	0,10	1,645
95%	0,05	1,96
99%	0,01	2,575



MARGEM DE ERRO

– Quando coletamos um conjunto de dados amostrais, podemos calcular a proporção amostral, a qual é tipicamente diferente da proporção populacional.

– A **margem de erro** (***E***) é a diferença máxima provável entre a proporção amostral observada e o verdadeiro valor da proporção populacional:

  – Isso ocorre quando dados de amostra aleatória simples são usados para estimar uma proporção populacional.
  – É também chamada de erro máximo da estimativa.
  – É encontrada pela multiplicação do valor crítico pelo desvio padrão das proporções amostrais.

MARGEM DE ERRO E INTERVALO DE CONFIANÇA

– Margem de erro para proporções é calculada por:

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

– Há uma probabilidade *α* de que a proporção amostral tenha erro maior do que *E*.

– Ou seja, $\hat{p}$ terá probabilidade de 1 – *α* de estar a: $z_{\alpha/2}\sqrt{pq/n}$ de *p*.

– Intervalo de confiança para proporção populacional é representado por: $\hat{p} - E < p < \hat{p} + E$

$$\hat{p} \pm E$$

$$(\hat{p} - E; \hat{p} + E)$$

EXEMPLO DE CÁLCULO

– Por exemplo, em 280 tentativas, houve 123 acertos:

  – n = 280

  – $\hat{p}$ = 123/280 = 0,439286

  – $\hat{q}$ = 1 – 0,439286 = 0,560714

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1{,}96\sqrt{\frac{(0{,}439286)(0{,}560714)}{280}} = 0{,}058133$$

$$\hat{p} - E < p < \hat{p} + E$$

$$0{,}439286 - 0{,}058133 < p < 0{,}439286 + 0{,}058133$$

$$0{,}381 < p < 0{,}497$$

– A taxa de sucesso é de 44%, com margem de erro de mais ou menos 6% e nível de confiança de 95% (geralmente resultados eleitorais omitem o nível de confiança).

COMO DEFINIR O TAMANHO AMOSTRAL?

– Utilizando a fórmula da margem de erro, chegamos a:

$$n = \frac{\left[z_{\alpha/2}\right]^2 \hat{p}\hat{q}}{E^2}$$

– Se não conhecemos qualquer estimativa $\hat{p}$:

$$n = \frac{\left[z_{\alpha/2}\right]^2 0{,}5 * 0{,}5}{E^2} = \frac{\left[z_{\alpha/2}\right]^2 0{,}25}{E^2}$$

– Se o tamanho amostral calculado não for um número inteiro, arredonde-o para o inteiro maior mais próximo.

– Quando a amostragem é sem reposição, a partir de uma população finita relativamente pequena, utilize:

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{\frac{N-n}{N-1}} \qquad n = \frac{\left[z_{\alpha/2}\right]^2 N\hat{p}\hat{q}}{\hat{p}\hat{q}\left[z_{\alpha/2}\right]^2 + (N-1)E^2}$$

EXEMPLO DE TAMANHO DA AMOSTRA

– Se margem de erro desejada igual a 5%, E=0,05.

– Se nível de confiança desejada é de 95%, $z_{\alpha/2}$=1,96.

– Assim:

$$n = \frac{\left[z_{\alpha/2}\right]^2 0{,}25}{E^2} = \frac{(1{,}96)^2 * 0{,}25}{(0{,}05)^2} = \frac{0{,}9604}{0{,}025} = 384{,}16 \approx 385$$

TAMANHO DA POPULAÇÃO

– Para o cálculo do tamanho da amostra, o tamanho da população é usado somente em casos em que fazemos amostragem sem reposição (dependente), a partir de uma população relativamente pequena.

– É prática comum considerarem-se os eventos como independentes (com reposição) quando pequenas amostras são retiradas de grandes populações.

– É raro selecionar o mesmo item duas vezes.

– Se o tamanho da amostra não é maior que 5% do tamanho da população, trate as seleções como sendo independentes.

– Isso é usado em pesquisas de opinião pública, quando há poucas entrevistas em uma população de milhões.

ESTIMAÇÃO DA MÉDIA POPULACIONAL: σ CONHECIDO

INTERVALO E NÍVEL DE CONFIANÇA, MARGEM DE ERRO

– O **intervalo de confiança** permite compreender melhor a precisão da estimativa da média amostral.

– Este intervalo está associado a um **nível de confiança**, o qual indica a taxa de sucesso do procedimento usado para construção do intervalo (confiabilidade).

– Diferença entre a média amostral e a média populacional é um erro.

– **Margem de erro** para a média, baseada em *σ* conhecido:

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

– Com isso, calculamos os limites do intervalo de confiança:

$$\bar{x} - E < \mu < \bar{x} + E \quad \text{ou} \quad \bar{x} \pm E \quad \text{ou} \quad (\bar{x} - E; \bar{x} + E)$$

TAMANHO AMOSTRAL PARA ESTIMAR MÉDIA μ

– Determinação do tamanho de amostra aleatória simples é importante, porque amostras grandes gastam tempo e dinheiro, e amostra pequenas levam a resultados imprecisos.

– Fórmula do tamanho amostral não depende do tamanho da população (*N*):

$$n = \left[\frac{z_{\alpha/2}\sigma}{E}\right]^2$$

– $z_{\alpha/2}$ = escore *z* crítico com base no nível de confiança.

– *E* = margem de erro desejada.

– σ = desvio padrão populacional.

– Caso de amostra sem reposição de população finita:

$$n = \frac{N\sigma^2\left(z_{\alpha/2}\right)^2}{(N-1)E^2 + \sigma^2\left(z_{\alpha/2}\right)^2}$$

LIDANDO COM σ DESCONHECIDO

– Geralmente o desvio padrão populacional é desconhecido.

– Use a **regra empírica da amplitude** para estimar o desvio padrão ($\sigma \approx$ amplitude/4).

  – Esse valor é maior ou igual ao real σ pelo menos 95% das vezes.

– Realize **estudo piloto**: comece processo de coleta da amostra e com base nos primeiros valores, calcule o desvio padrão amostral (s) e use-o no lugar de σ.

  – Esse valor pode ser melhorado à medida que mais dados são obtidos.

– Estime valor de σ com resultados de **estudos anteriores**.

– Ao calcular n, **erros devem ser conservadores**, no sentido de aumentar tamanho amostral em vez de diminuir.

ESTIMAÇÃO DA MÉDIA POPULACIONAL: σ DESCONHECIDO

MELHOR ESTIMATIVA DA MÉDIA POPULACIONAL

– A média amostral $\bar{x}$ continua sendo a melhor estimativa pontual da média populacional μ.

– Se σ não é conhecido, mas requisitos são satisfeitos, usamos **distribuição *t* de Student** (em vez de distribuição normal).

– O valor de σ é estimado com o valor do desvio padrão amostral (s), mas isso introduz fonte de não-confiabilidade, principalmente quando amostras são pequenas.

– Isso é compensado fazendo o intervalo de confiança um pouco mais largo, com os valores críticos $t_{\alpha/2}$ que são maiores do que os valores críticos $z_{\alpha/2}$.

MARGEM DE ERRO E INTERVALO DE CONFIANÇA

– Para calcular margem de erro *E* para estimativa de μ com σ desconhecido, onde $t_{\alpha/2}$ tem n–1 graus de liberdade:

$$E = t_{\alpha/2} \frac{s}{\sqrt{n}}$$

– Intervalo de confiança para estimativa de μ com σ desconhecido:

$$\bar{x} - E < \mu < \bar{x} + E$$

DISTRIBUIÇÃO *t* DE STUDENT PARA *n*=3 E *n*=12

– Distribuição *t* de Student tem a mesma forma geral da distribuição normal padrão, mas reflete a maior variabilidade que se espera com amostras pequenas.