MAT02025 - Amostragem 1

AAS: intervalos de confiança para uma proporção

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2022

IC aproximado para P

IC aproximado para P

- Da expressão para a variância estimada de P, uma forma da aproximação normal para os limites de confiança de P é

$$p \pm \left[z\sqrt{1-f}\sqrt{pq/(n-1)} + \frac{1}{2n}\right],$$

em que $f = n/N$, z é o desvio normal correspondente à probabilidade de confiança.

IC aproximado para P

- O uso do termo mais familiar $\sqrt{pq/n}$ raramente faz uma diferença apreciável.
- O último termo à direita $(1/2n)$ é uma **correção para continuidade**.
  - Isso produz apenas uma ligeira melhora na aproximação.
  - No entanto, sem a correção, a aproximação normal geralmente fornece um intervalo de confiança muito estreito.

IC aproximado para P

- O erro na aproximação normal depende de todas as quantidades n, p, N e α ($1-\alpha$ é coeficiente de confiança do intervalo).
- A quantidade à qual o erro é mais sensível é np ou mais especificamente o número observado na classe menor.
- A Tabela a seguir fornece regras de trabalho para decidir quando a aproximação normal pode ser usada.

p	np = número observado na classe menor	n = tamanho da amostra
0,5	15	30
0,4	20	50
0,3	24	80
0,2	40	200
0,1	60	600
0,05	70	1400
≈ 0	80	∞

IC aproximado para P

- As regras apresentadas na tabela acima são construídas de modo que, com limites de confiança de 95%, a frequência real com a qual os limites falham em incluir P não seja maior que 5,5%.
- Além disso, a probabilidade de que o limite superior esteja abaixo de P está entre 2,5 e 3,5%, e a probabilidade de que o limite inferior exceda P está entre 2,5 e 1,5%.

IC exato para P

IC exato para P

- Os limites de confiança também podem ser obtidos com base na **distribuição hipergeométrica** exata do número de unidades na amostra com o atributo.
- O método exato é conceitualmente simples, mas computacionalmente complexo.
- Seja $a = \sum_{i=1}^{n} Y_i$ o número de unidades com o atributo (pertencentes a classe C) na amostra.

IC exato para P

- Para um intervalo de confiança de 100 (1 - α)% desejado para o número A de unidades na população com o atributo, um limite superior $\widehat{A}_S$ é determinado como de unidades na população com o atributo dando probabilidade α_1 de obter a ou menos unidades com o atributo na amostra, em que α_1 é aproximadamente igual a metade do α desejado.
- Ou seja, $\widehat{A}_S$ satisfaz

$$\Pr(X \leq a) = \sum_{j=0}^{a} \Pr(j, n-j | \widehat{A}_S, N - \widehat{A}_S) = \sum_{j=0}^{a} \binom{\widehat{A}_S}{j} \binom{N - \widehat{A}_S}{n-j} \Big/ \binom{N}{n} = \alpha_1.$$

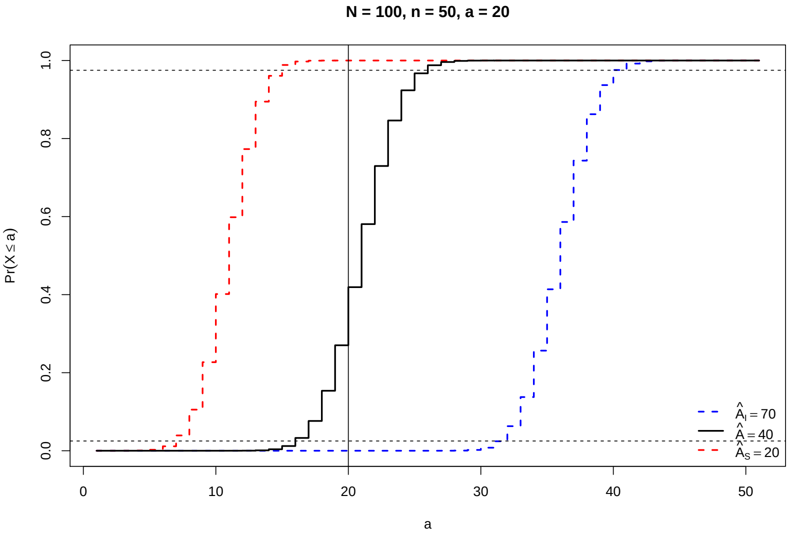
IC exato para P

- O limite inferior $\widehat{A}_I$ é o número de unidades na população com o atributo dando probabilidade α_2 de se obter a ou mais unidades com o atributo na amostra, em que α_2 é aproximadamente igual a metade do α desejado.
- Ou seja, $\widehat{A}_I$ satisfaz

$$\Pr(X \geq a) = \sum_{j=a}^{n} \Pr(j, n-j|\widehat{A}_I, N-\widehat{A}_I) = \sum_{j=a}^{n} \binom{\widehat{A}_I}{j}\binom{N-\widehat{A}_I}{n-j} \Big/ \binom{N}{n} = \alpha_2.$$

- Os limites de confiança para P são então determinados, dividindo-se os limites achados para A por N, ou seja: $\widehat{P}_I = \widehat{A}_I/N$ e $\widehat{P}_S = \widehat{A}_S/N$.

IC exato para P



Algoritmo para obter o IC extato para P

Procuramos os limites de confiança ótimos $(\widehat{A}_I, \widehat{A}_S)$ que atendem aos requisitos definidos nas equações acima.

- Dada a população total conhecida N, o tamanho da amostra n e o número de "sucessos" na amostra a, podemos definir alguns limites de viabilidade para A:
  - Naturalmente, o menor valor possível é o número observado de sucessos $A_{min} = a$
  - O maior valor possível é igual ao número total N menos as observações na amostra que pertencem a classe C', ou seja, $A_{max} = N - (n - a)$.

Algoritmo para obter o IC extato para P

- Limite superior $\widehat{A}_S$:
  - Comece com o maior valor possível para A, ou seja, $A_{max} = N - (n - a)$;
  - Então, diminua incrementalmente enquanto o $\Pr(X \leq a) < \alpha/2$, de modo que encontremos o maior valor possível que ainda satisfaz a equação.
- Limite inferior $\widehat{A}_I$:
  - Comece com o menor valor possível para A, ou seja, $A_{min} = a$;
  - Reescrever $\Pr(X \geq a) = 1 - \Pr(X \leq a) = \alpha/2 \Leftrightarrow \Pr(X \leq a) = 1 - \alpha/2$;
  - Então, aumente incrementalmente enquanto $\Pr(X \leq a) \geq 1 - \alpha/2$, de modo que encontremos o menor valor possível que ainda preenche a equação.

Exemplo

Exemplo

- Em um levantamento por amostragem, utilizando amostragem aleatória simples sem reposição, de tamanho $n = 100$, de uma população de tamanho 500, foi observado que 37 indivíduos são favoráveis a adoção de uma certa política pública.
  - Os demais são contrários ou não sabem opinar.
- Os limites de confiança de 95% para a proporção e para o número total de unidades da categoria C na população podem ser obtidos utilizando a aproximação normal e a distribuição hipergeométrica.

Exemplo

Aproximação normal

- O erro padrão estimado de p é

$$\sqrt{1-f}\sqrt{pq/(n-1)} = \sqrt{0,8}\sqrt{(0,37)(0,63)/99} = 0,0434.$$

- A correção de continuidade, $1/2n$, é igual a 0,005. Portanto, os limites de 95% para P podem ser estimados como

$$0,37 \pm (1,96 \times 0,0434 + 0,005) = 0,37 \pm 0,090 = (0,280; 0,460).$$

- Para achar os limites para o número total de unidades da população que pertencem à categoria C, multiplicamos os valores acima por N e obtemos 140 e 230, respectivamente.

Exemplo

Distribuição hipergeométrica

```
# install.packages("samplingbook")
library(samplingbook)

Sprop(m = 37, n = 100, N = 500, level = 0.95)
```

```
##
## Sprop object: Sample proportion estimate
## With finite population correction: N = 500
##
## Proportion estimate:  0.37
## Standard error:  0.0434
##
## 95% approximate confidence interval:
##  proportion: [0.2849,0.4551]
##  number in population: [143,227]
## 95% exact hypergeometric confidence interval:
##  proportion: [0.284,0.464]
##  number in population: [142,232]
```

Considerações finais sobre a aproximação normal

- Na maioria dos cenários, essa estratégia (intervalo de confiança utilizando a aproximação normal) resulta em propriedades satisfatórias.
- No entanto, se p estiver próximo de 0 ou 1, é recomendado usar o intervalo de confiança exato com base na distribuição hipergeométrica[1].
- O intervalo aproximado tem uma **probabilidade de cobertura** tão baixa quanto n/N para qualquer α. Portanto, não há garantia de que o intervalo capture o verdadeiro A com o nível de confiança desejado se a amostra for muito menor do que a população[2].

[1] Kauermann, Goeran, and Helmut Kuechenhoff. 2010. *Stichproben: Methoden Und Praktische Umsetzung Mit R*. Springer-Verlag.

[2] Wang, Weizhen. 2015. Exact Optimal Confidence Intervals for Hypergeometric Parameters. *Journal of the American Statistical Association* 110 (512): 1491–9.

Considerações finais sobre a aproximação normal

- Ainda, com p e n pequenos, o IC aproximado pode produzir limites inferiores a 0:

```
Sprop(m = 2, n = 100, N = 500, level = 0.95)
```

```
##
## Sprop object: Sample proportion estimate
## With finite population correction: N = 500
##
## Proportion estimate:  0.02
## Standard error:  0.0126
##
## 95% approximate confidence interval:
##  proportion: [-0.0047,0.0447]
##  number in population: [-2,22]
## 95% exact hypergeometric confidence interval:
##  proportion: [0.004,0.066]
##  number in population: [2,33]
```

Para casa

- Revisar os tópicos discutidos nesta aula.
- Implementar o IC para P utilizando a distribuição binomial como aproximação da distribuição hipergeométrica.

Próxima aula

- Classificação em mais de duas categorias.

Por hoje é só!

Bons estudos!