MAT02025 - Amostragem 1

AAS: distribuição das estimativas de P e 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, 2023

Influência de P no erro padrão

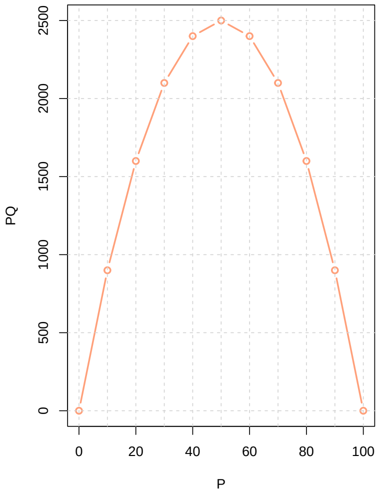
Influência de P no erro padrão

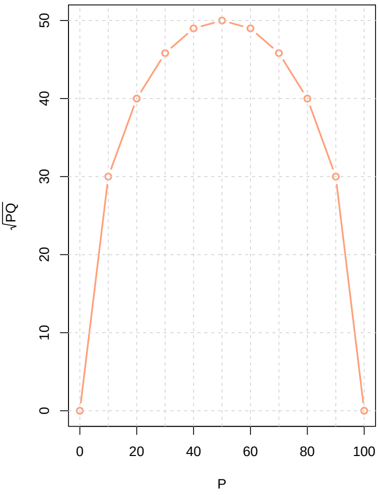
- A equação (2), da aula passada, mostra como a variância da porcentagem estimada muda com P (a **porcentagem da população na categoria C**), para n e N fixos. Se a cpf for ignorada, temos

$$\mathrm{Var}\,(p) = \frac{PQ}{n}.$$

- A função PQ e sua raiz quadrada são mostradas a seguir.
  - Essas funções podem ser consideradas como variância e desvio padrão, respectivamente, para uma amostra de tamanho 1.

Influência de P no erro padrão



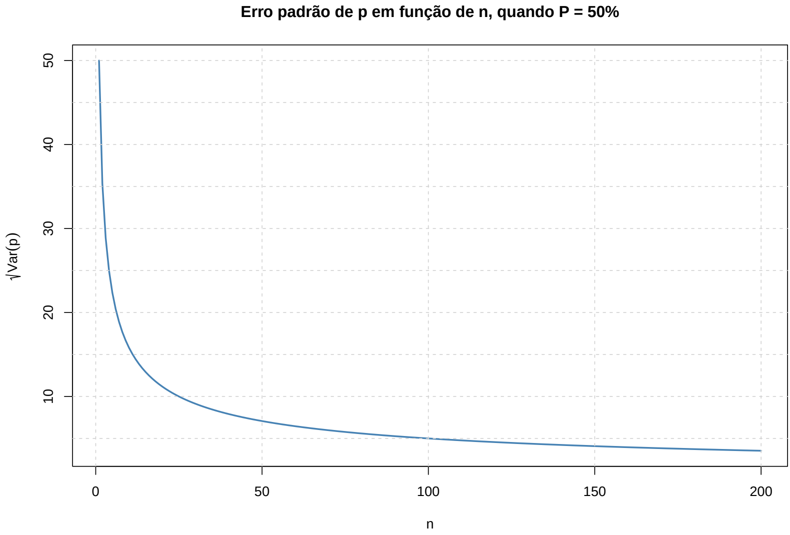


Influência de P no erro padrão

Observações

- As funções têm seus maiores valores quando a população é dividida igualmente entre as duas classes e são simétricas em relação a este ponto.
- O erro padrão de p muda relativamente pouco quando P está entre 30 e 70%.
- No valor máximo de $\sqrt{PQ}$, 50, um tamanho de amostra de 100 é necessário para reduzir o erro padrão da estimativa para 5%.
- Para atingir um erro padrão de 1%, é necessário um tamanho de amostra de 2500.

Influência de P no erro padrão



Influência de P no erro padrão

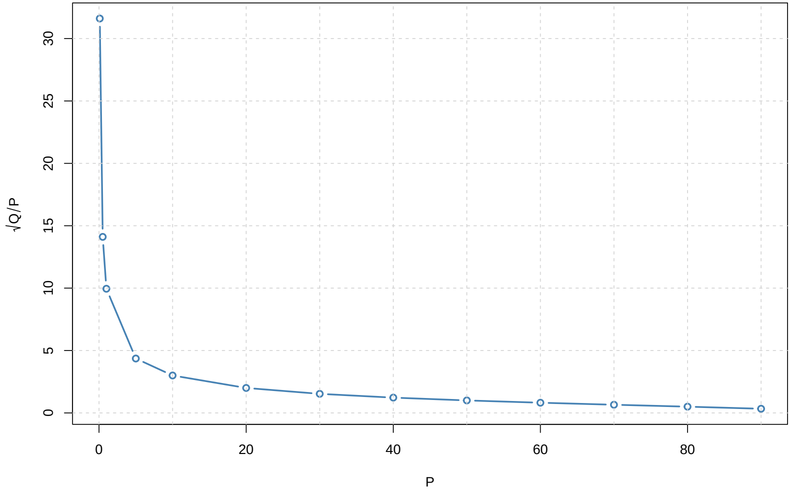
- Esta abordagem não é apropriada quando o interesse reside no **número total** de unidades da população que estão na classe C.
- Nesse caso, é mais natural perguntar: a estimativa provavelmente está correta dentro de, digamos, 7% do verdadeiro total?
- Assim, tendemos a pensar no erro padrão expresso como uma fração ou porcentagem do valor verdadeiro, NP. A fração é

$$\frac{\sigma_{N_p}}{NP} = \frac{N\sqrt{PQ}}{\sqrt{n}NP}\sqrt{\frac{N-n}{N-1}} = \frac{1}{\sqrt{n}}\sqrt{\frac{Q}{P}}\sqrt{\frac{N-n}{N-1}}.$$

Influência de P no erro padrão

- Essa quantidade é chamada de **coeficiente de variação** da estimativa.
- Se a cpf for ignorada, o coeficiente é $\sqrt{Q/nP}$.
- A razão $\sqrt{Q/P}$, que pode ser considerada o coeficiente de variação para uma amostra de tamanho 1, é mostrada a seguir.

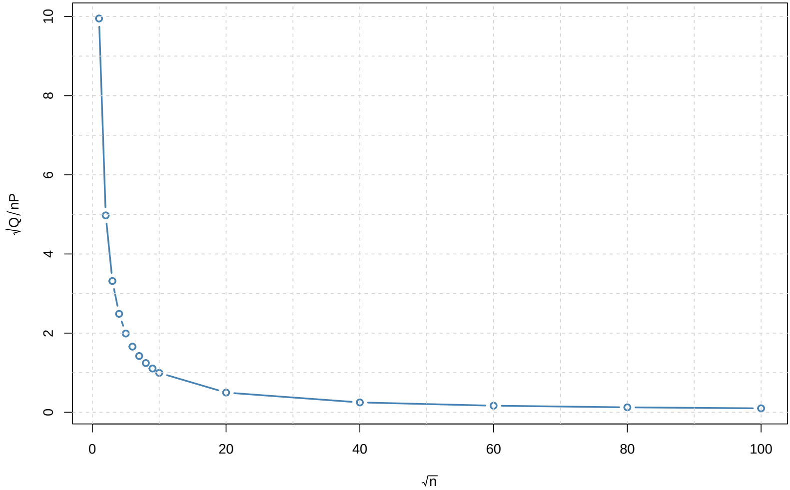
Influência de P no erro padrão



Influência de P no erro padrão

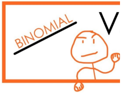
- Para um tamanho de amostra fixo, o coeficiente de variação do total estimado na classe C diminui continuamente à medida que a porcentagem verdadeira em C aumenta.
- O coeficiente é alto quando P é menor que 5%.
- Amostras muito grandes são necessárias para estimativas precisas do número total que possui qualquer atributo raro na população.
- Para $P = 1\%$, devemos ter $\sqrt{n} = 99$ para reduzir o coeficiente de variação da estimativa para 0,1 ou 10%.
  - Isso dá um tamanho de amostra de 9801.
  - A amostragem aleatória simples, ou qualquer método de amostragem que seja adaptado para propósitos gerais, é um método caro de estimar o número total de unidades de um tipo escasso.

Influência de P no erro padrão



Distribuição binomial

Distribuição binomial



- Como a população é de um tipo particularmente simples, em que os Y_i são 1 ou 0, podemos encontrar a **distribuição de frequência real** da estimativa p e não apenas sua **média** e **variância**.

Distribuição binomial

- A população contém A unidades que estão na classe C e $(N-A)$ unidades em C', em que $P = A/N$.
- Se a primeira unidade sorteada estiver em C, permanecerão na população $(A-1)$ unidades em C e $N-A$ em C'.
- Assim, a proporção de unidades em C, após o primeiro sorteio, muda ligeiramente para $(A-1)/(N-1)$.
- Alternativamente, se a primeira unidade selecionada estiver em C', a proporção em C muda para $A/(N-1)$.

Distribuição binomial

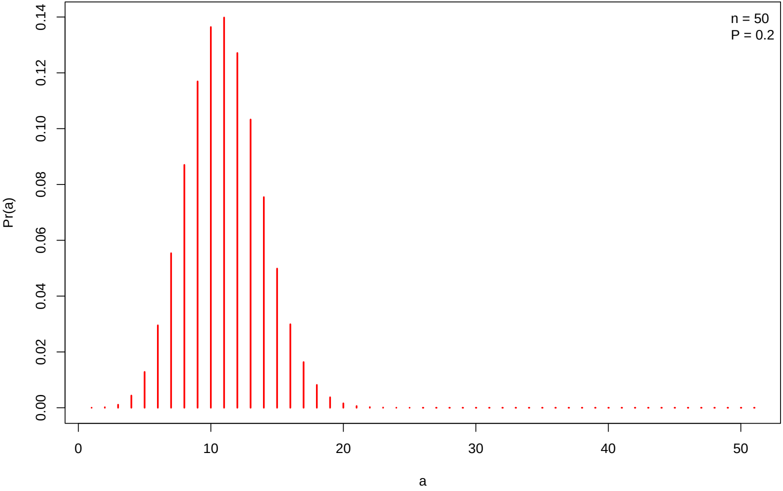
- Na amostragem sem reposição, a proporção continua mudando dessa forma ao longo do sorteio das unidades.
- Na presente seção, essas variações são ignoradas, ou seja, P é considerado **constante**.
- Isso equivale a supor que A e $(N - A)$ **são ambos grandes** em relação ao tamanho da amostra n (ou que a amostragem é feita com reposição).

Distribuição binomial

- Com essa suposição, o processo de sorteio da amostra consiste em uma série de n tentativas, em cada uma das quais a probabilidade de que a unidade selecionada esteja em C é P.
- Esta situação dá origem à **distribuição de frequência binomial** para o número de unidades em C na amostra.
- A probabilidade de que a amostra contenha a unidades em C é

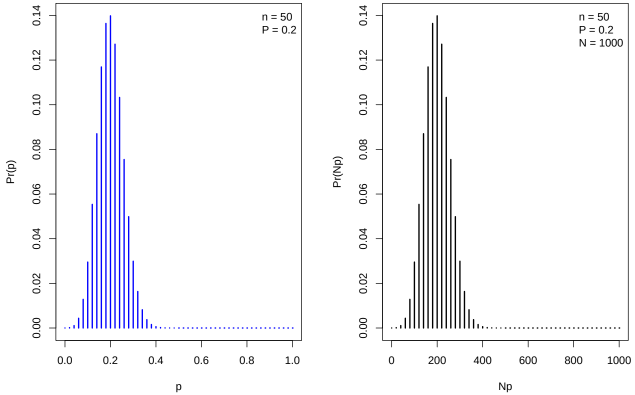
$$\Pr(a) = \frac{n!}{a!(n-a)!} P^a Q^{n-a}, \quad a = 0, 1, \ldots, n.$$

Distribuição binomial



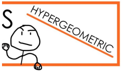
Distribuição binomial

- A partir dessa expressão, podemos tabular a distribuição de frequência de a, de $p = a/n$ ou do total estimado Np.



Distribuição hipergeométrica

Distribuição hipergeométrica



- A distribuição de p pode ser encontrada sem a suposição de que a população seja grande em relação à amostra.
- O número de unidades nas duas classes C e C' na população são A e A', respectivamente.
- Vamos calcular a probabilidade de que os números correspondentes na amostra sejam a e a', em que

$$a + a' = n, \quad A + A' = N.$$

Distribuição hipergeométrica

- Na amostragem aleatória simples, cada uma das $\binom{N}{n}$ diferentes seleções de n unidades de N tem uma chance igual de ser sorteada.
- Para encontrar a probabilidade desejada, contamos quantas dessas amostras contêm exatamente a unidades de C e a' de C'.
- O número de seleções diferentes de a unidades entre A que está em C é $\binom{A}{a}$, enquanto o número de seleções diferentes de a' entre A' é $\binom{A'}{a'}$.
- Cada seleção do primeiro tipo pode ser combinada com qualquer uma do segundo para dar uma amostra diferente do tipo necessário.
- O número total de amostras do tipo necessário é, portanto,

$$\binom{A}{a} \times \binom{A'}{a'}.$$

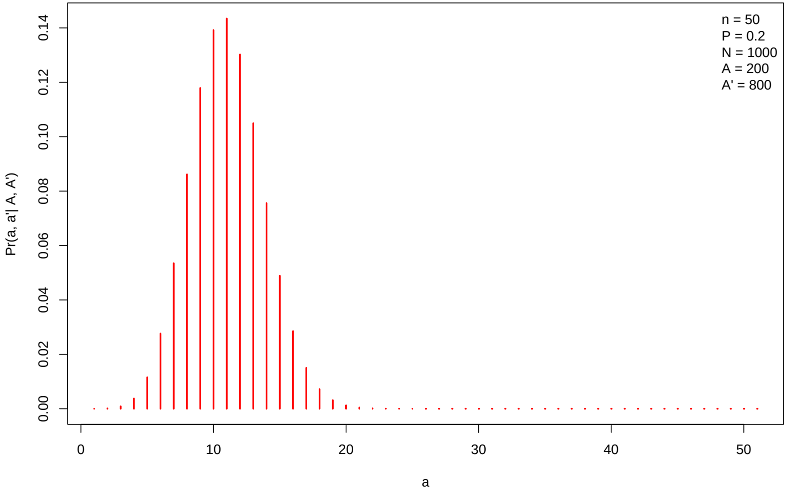
Distribuição hipergeométrica

- Portanto, se uma amostra aleatória simples de tamanho n for sorteada, a probabilidade de que seja do tipo necessário é

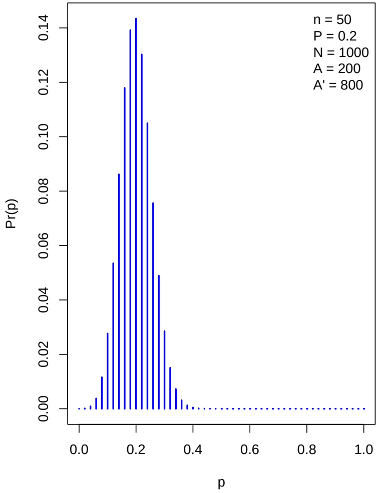
$$\Pr(a, a'|A, A') = \frac{\binom{A}{a}\binom{A'}{a'}}{\binom{N}{n}}$$

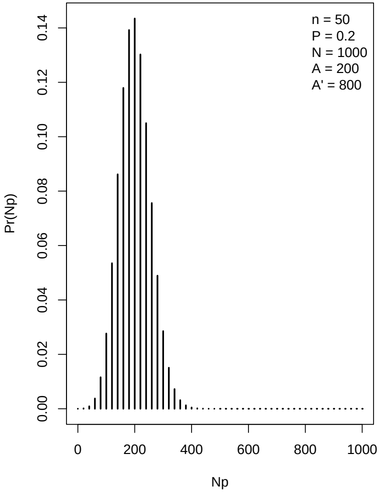
- Esta é a distribuição de frequência de a ou np, da qual a distribuição de p é imediatamente derivada.
- A distribuição é chamada de **distribuição hipergeométrica**.

Distribuição hipergeométrica



Distribuição hipergeométrica



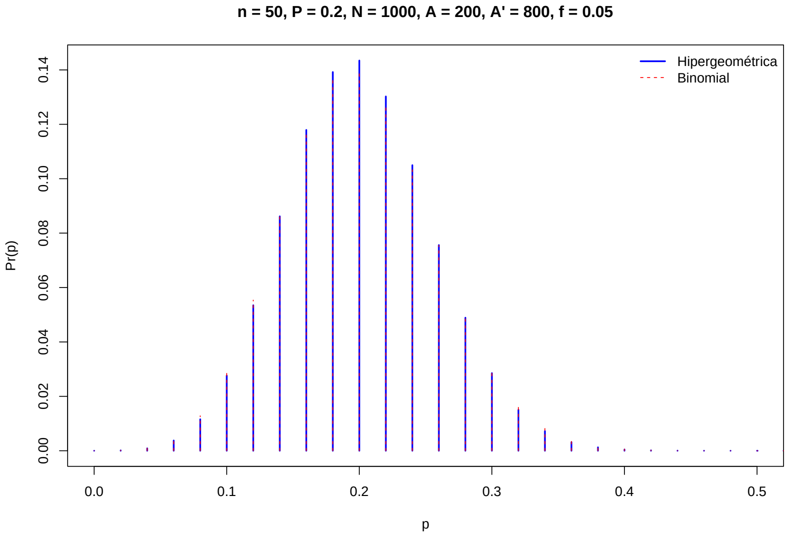


A binomial é uma boa aproximação para a hipergeométrica?

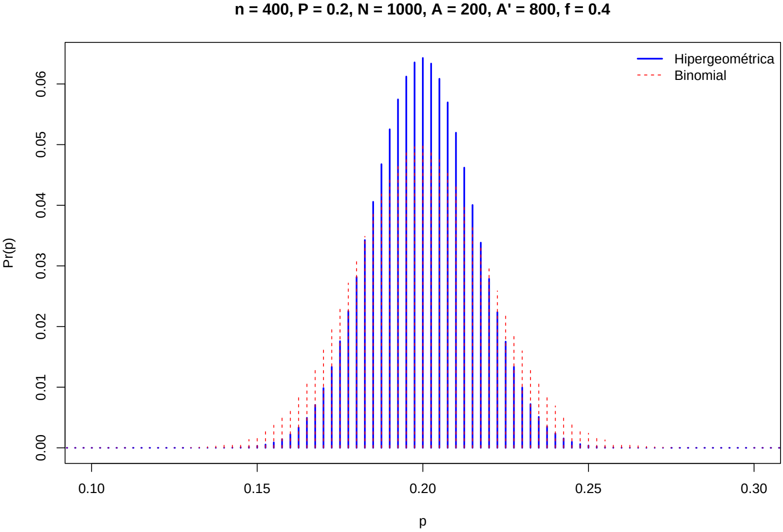
Qualidade da aproximação

- **Relembrando:**
  - P é considerado **constante**.
    - Isso equivale a supor que A e $N - A$ **são ambos grandes** em relação ao tamanho da amostra n **(fração de amostragem é pequena)**.

Qualidade da aproximação



Qualidade da aproximação



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 $\sqrt{pq/n}$ não apresenta uma diferença notável.
- O último termo à direita $(1/2n)$ é uma **correção de 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**.
- **Para pensar:** intervalos com menor amplitude representam menor incerteza com respeito a estimativa do parâmetro de interesse. Por outro lado, espera-se que $100 \times (1 - \alpha)\%$ dos intervalos de confiança de $100 \times (1 - \alpha)\%$ contenham o verdadeiro parâmetro. Intervalos "encolhidos" podem não garantir que a **taxa de cobertura** dos ICs seja próxima à respectiva **taxa nominal**.

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 menor classe.
- A tabela a seguir fornece **regras de trabalho** para decidir quando a aproximação normal pode ser usada.

p	np = número observado na menor classe	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%.
  - Ou seja, a taxa de cobertura dos ICs de 95% com base na aproximação normal (dadas as condições da tabela) não deve ser inferior a 94,5%.
- 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 para A

- Para obter os limites de confiança para o parâmetro A, número de unidades que pertencem a classe C na população, multiplicamos por N os limites inferior e superior do intervalo de confiança para P.
  - $\widehat{A}_I = N\widehat{P}_I$ e $\widehat{A}_S = N\widehat{P}_S$.

IC exato para A

IC exato para A

- Os limites de confiança também podem ser obtidos com base na **distribuição hipergeométrica**.
  - **Lembrando:** esta é a **distirbuição exata** do número de unidades na amostra que pertencem a classe C, a.
- O **método exato** de obtenção do intervalo de confiança para A é **conceitualmente simples**, mas **computacionalmente complexo**.

IC exato para A

- Seja $a = \sum_{i=1}^{n} Y_i$ o número de unidades pertencentes a classe C na amostra.
- Para um intervalo de confiança de $100 \times (1 - \alpha)\%$ desejado para o número A, um limite superior $\widehat{A}_S$ é determinado como o número de unidades na população que pertencem a classe C que fornece probabilidade α_1 de obter a ou menos unidades que pertencem a classe C na amostra, em que α_1 é aproximadamente igual a metade do α desejado.

IC exato para A

- Ou seja, $\widehat{A}_S$ satisfaz

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

IC exato para A

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

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

- 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 $\widehat{P}_S = \widehat{A}_S / N$.

Algoritmo para obter o IC extato para A

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

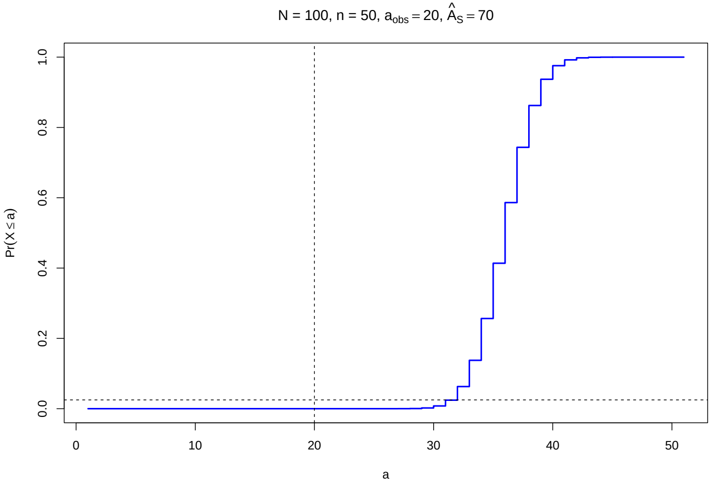
Algoritmo para obter o IC extato para A

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.

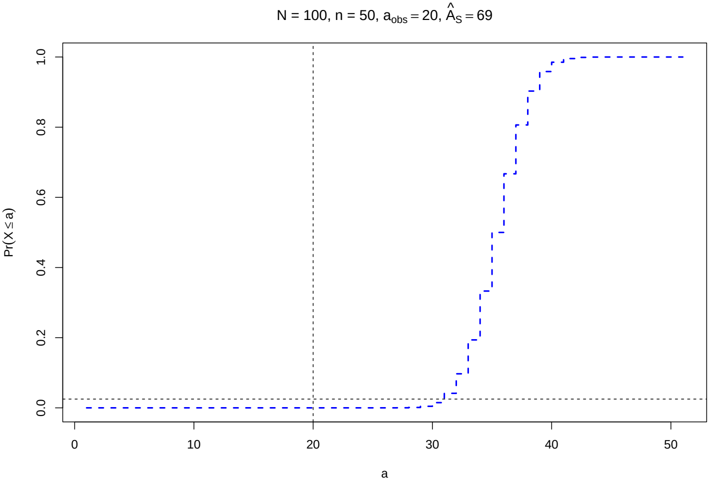
Algoritmo para obter o IC extato para A

- Suponha $\alpha = 0,05$. $\Pr(X \leq a | \widehat{A}_S = 70) < 0,025$.

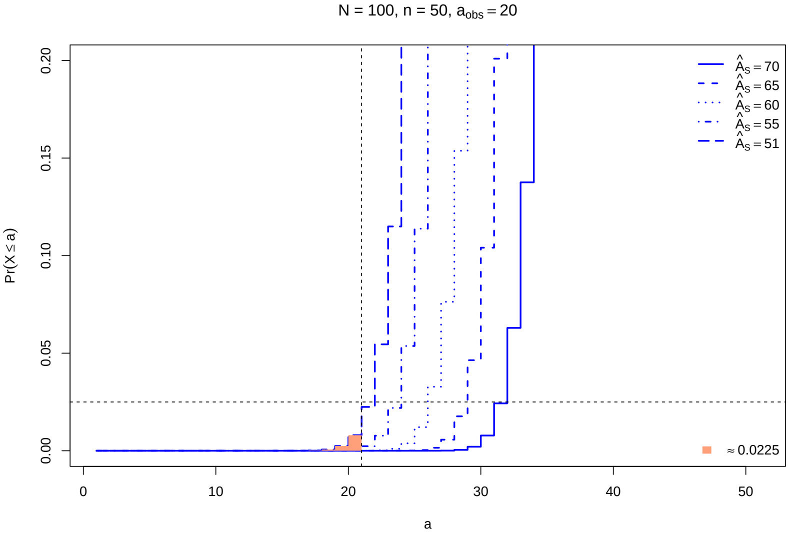


Algoritmo para obter o IC extato para A

- Supondo $\alpha = 0,05$. $\Pr(X \leq a|\widehat{A}_S = 69) < 0,025$.



Algoritmo para obter o IC extato para A

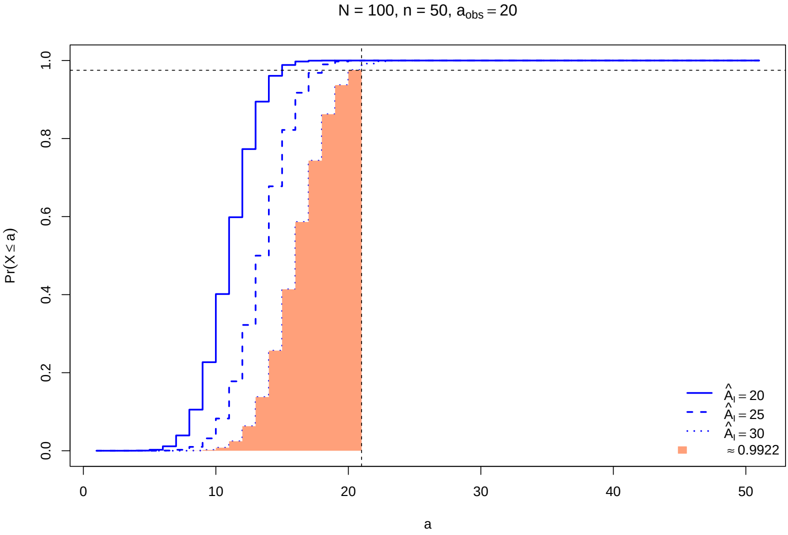


Algoritmo para obter o IC extato para A

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.

Algoritmo para obter o IC extato para A



Exemplo

Exemplo

- Em um **levantamento por amostragem**, utilizando **amostragem aleatória simples sem reposição**, de tamanho $n = 100$, de uma população de tamanho $N = 500$, foi observado que $a = 37$ indivíduos são favoráveis a adoção de uma certa política pública (por exemplo, a adoção da **semana de 4 dias de trabalho**).
  - Os demais são contrários ou não sabem opinar.
- Os limites de confiança de 95% para a **proporção** e para o **número total de unidades** que pertencem a classe C (**favoráveis a adoção da política pública**) na população podem ser obtidos utilizando a aproximação normal e a distribuição hipergeométrica.

Exemplo

IC para P utilizando 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

$$\begin{aligned} IC(P;95\%) &= 0,37 \pm (1,96 \times 0,0434 + 0,005) \\ &= 0,37 \pm 0,090 = (0,280;0,460). \end{aligned}$$

Exemplo

IC para A utilizando aproximação normal

- Para achar os limites para o número total A de unidades da população que pertencem à categoria C, multiplicamos os valores acima por N:

$$
\begin{aligned}
IC(A; 95\%) &= (500 \times 0,280; 500 \times 0,460) \\
&= (140; 230)
\end{aligned}
$$

Exemplo

IC para A e P extato

- Como visto anteriormente, o intervalo de confiança exato é baseado na **distribuição hipergeométrica** (distribuição exata de a) e requer a avaliação da distribuição acumalada.
- O pacote `samplingbook` do `R` possui uma função (`Sprop`) que facilita o trabalho do profissonal de estatística.

Exemplo

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

# ?Sprop
Sprop(m = 37, # m = a
      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

Considerações finais

Sobre a aproximação normal

- Na maioria dos cenários, a construção do 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].
- Ainda, com p e n pequenos, o IC aproximado pode produzir **limites inferiores** menores que 0.

[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

```r
Sprop(m = 2, n = 30, # n e p pequenos
      N = 500, level = 0.95)
```

```
##
## Sprop object: Sample proportion estimate
## With finite population correction: N = 500
##
## Proportion estimate:  0.0667
## Standard error:  0.0449
##
## 95% approximate confidence interval:
##  proportion: [-0.0214,0.1547]
##  number in population: [-10,77]
## 95% exact hypergeometric confidence interval:
##  proportion: [0.008,0.218]
##  number in population: [4,109]
```

Para casa

- Revisar os tópicos discutidos nesta aula.
- Como podemos obter as probabilidades referentes a distribuições binomial e hipergeométrica?
- Rodar a simução de Monte Carlo para avaliar as taxas de cobertura dos ICs para P considerando diferentes tamanhos de população, amostra, valores de P e de α.
- Implementar o IC para P utilizando a distribuição binomial como aproximação da distribuição hipergeométrica.

Próxima aula

- Porporações para classificações em mais de duas categorias.

Por hoje é só!

Bons estudos!