MAT02025 - Amostragem 1

AAS: distribuição das estimativas de P

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, 2021

Influência de P no erro padrão

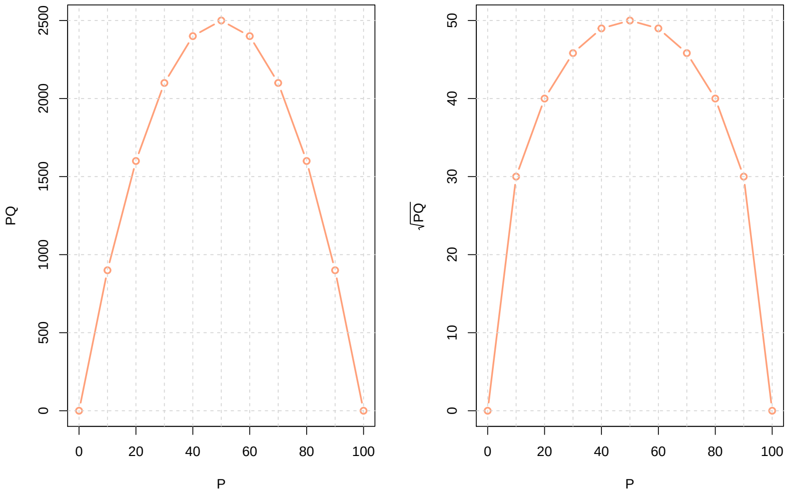
Influência de P no erro padrão

- 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

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

- 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

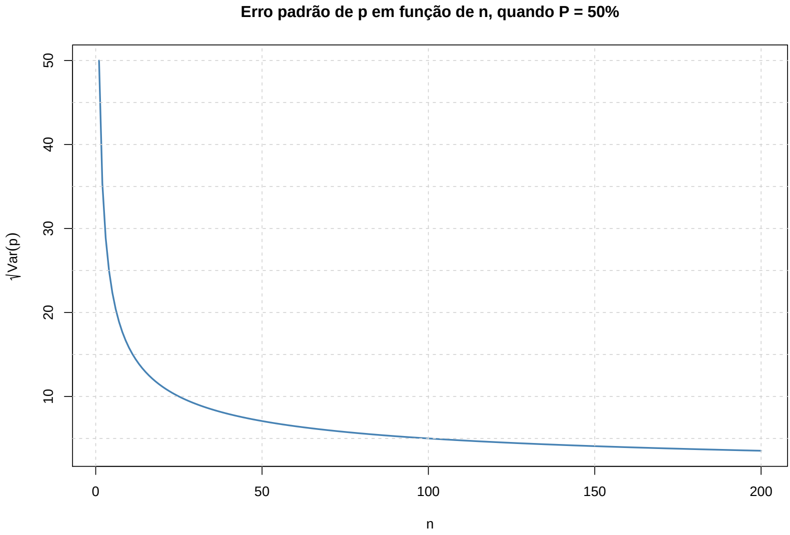


Influência de P no erro padrão

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

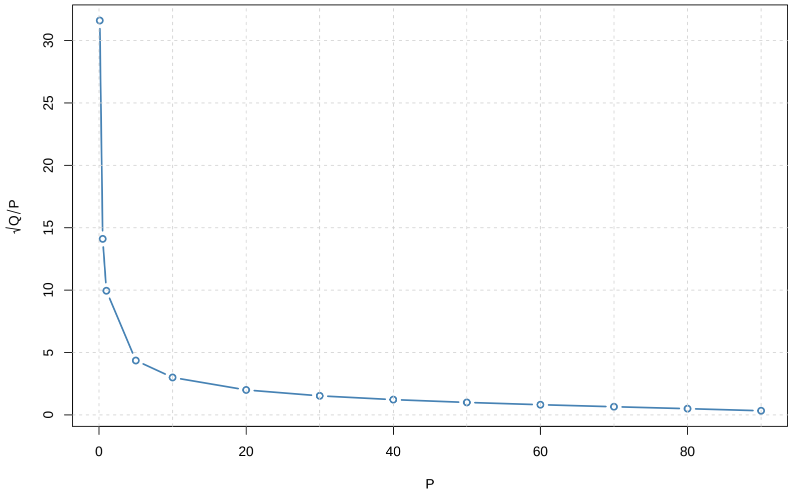
- Esta abordagem não é apropriada quando o interesse reside no **número total** de unidades da população que estão na classe C.
- Nesse caso, é mais natural perguntar: a estimativa provavelmente está correta dentro de, digamos, 7% do verdadeiro total?
- Assim, tendemos a pensar no erro padrão expresso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

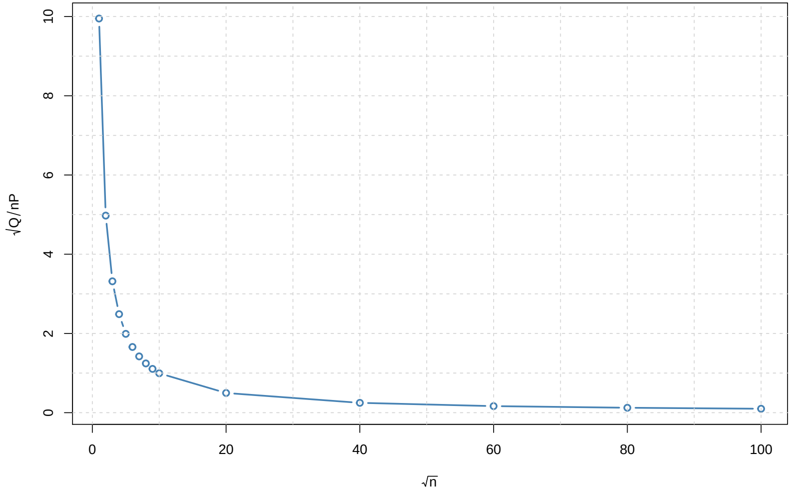
Influência de P no erro padrão



Influência de P no erro padrão

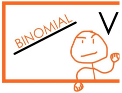
- Para um tamanho de amostra fixo, o coeficiente de variação do total estimado na classe C diminui continuamente à medida que a porcentagem verdadeira em C aumenta.
- O coeficiente é alto quando P é menor que 5%.
- Amostras muito grandes são necessárias para estimativas precisas do número total que possui 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 unidades $A - 1$ 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

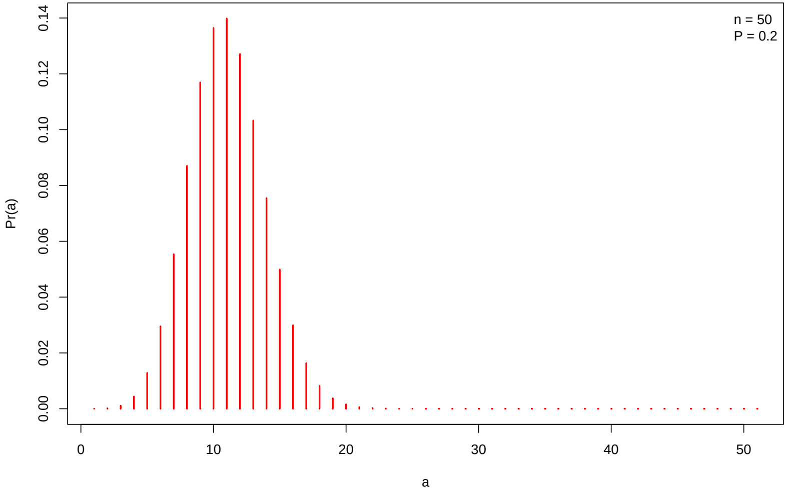
- Na amostragem sem reposição, a proporção continua mudando dessa forma ao longo do sorteio.
- Na presente seção, essas variações são ignoradas, ou seja, P é considerado **constante**.
- Isso equivale a supor que A e $N - A$ **são ambos grandes** em relação ao tamanho da amostra n (ou que a amostragem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

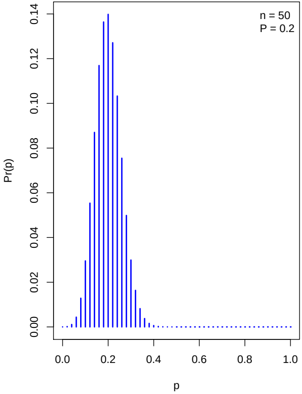
$$\Pr(a) = \frac{n!}{a!(n-a)!} P^a Q^{n-a}, \quad a = 0, 1, \ldots, n.$$

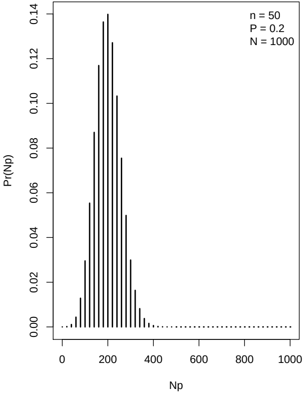
Distribuição binomial



Distribuição binomial

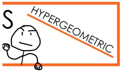
- A partir dessa expressão, podemos tabular a distribuição de frequência de a, de $p = a/n$ ou do total estimado Np.





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,

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

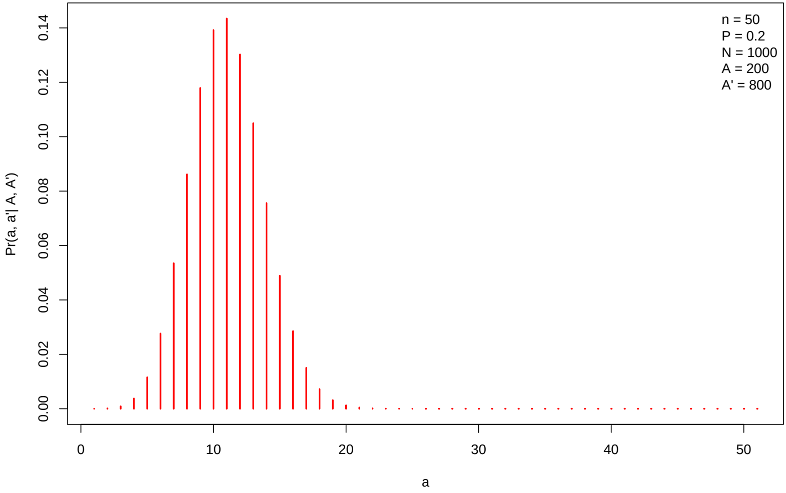
Distribuição hipergeométrica

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

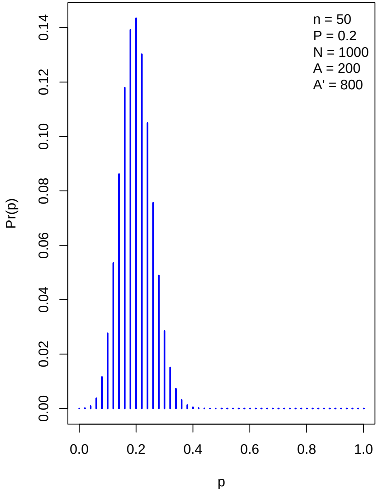
$$\Pr(a, a' | A, A') = \frac{\binom{A}{a}\binom{A'}{a'}}{\binom{N}{n}}$$

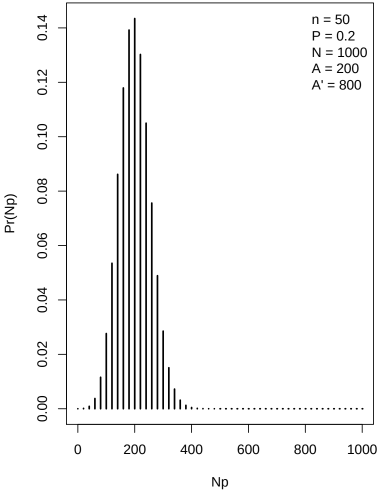
- Esta é a distribuição de frequência de a ou np, da qual a distribuição de p é imediatamente derivada.
- A distribuição é chamada de **distribuição hipergeométrica**.

Distribuição hipergeométrica



Distribuição hipergeométrica



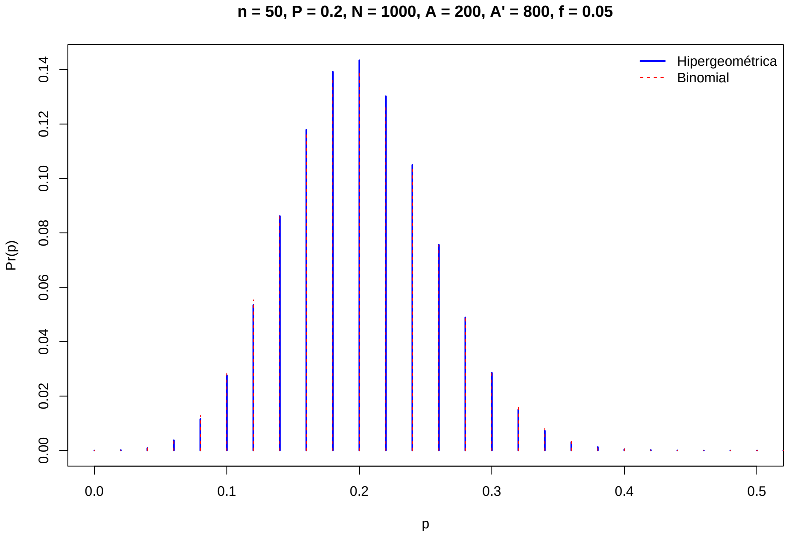


A binomial é uma boa aproximação para a hipergeométrica?

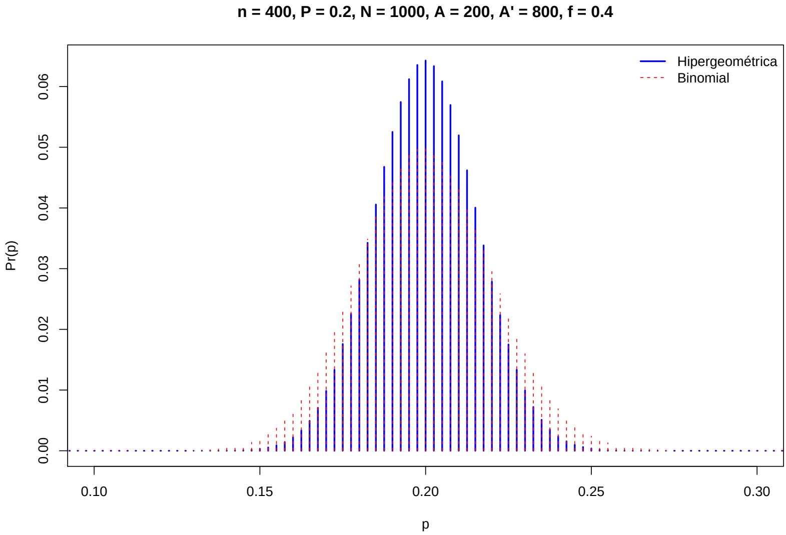
Qualidade da aproximação

- **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



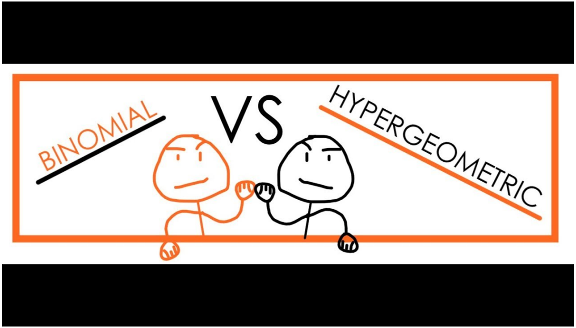
Para casa

Para casa

- Revisar os tópicos discutidos nesta aula.
- Como podemos obter as probabilidades referentes a distribuições binomial e hipergeométrica?

Próxima aula

- Intervalos de confiança para P:
  - Aproximados e exato (a real batalha).



Por hoje é só!

Bons estudos!