

# MAT02036 - Amostragem 2

## Aula 09 - Amostragem Estratificada - Estimação de Proporções

Markus Stein

Departamento de Estatística, IME/UFRGS

2022/2

## *Housekeeping*

- Aproveitem o momento presencial para tirar dúvidas
- Se estivéssemos no ensino remoto ou à distância
  - vocês poderiam estar somente ouvindo, sem interação
  - ou assistindo vídeos e material em outro momento
- Depois das aulas, rever material da aula passada
  - fazer exercícios
  - se preparar para a próxima aula

# Aula passada

## Exercícios para entregar 1

- Exercício 4.1 (Bolfarine e Bussab)
- Exercício 11.10 (Amostragem: Teoria e Prática Usando o R)
- Exercício 11.7 (Amostragem: Teoria e Prática Usando o R)

Dúvidas?

# Estimação de Proporções na AES

# Estimação de Proporções na AES

Um caso especial do que já vimos ocorre quando a variável  $y$  indicadora de uma unidade populacional  $i$  tem ou não uma determinada **característica** ou **atributo de interesse**, ou pertence a um determinado grupo especificado de unidades da população:

- Migrantes entre os habitantes de determinada região.
- Estabelecimentos agropecuários de produção leiteira numa localidade.
- Estudantes do sexo feminino em escolas.
- Sondagens eleitorais, parcela dos eleitores pretende votar em determinado candidato.

Definimos a variável  $y$ , para cada unidade  $i$  da população, assumindo um de dois valores possíveis:

$$y_i = I(i \in A) = \begin{cases} 1, & \text{se a unidade } i \text{ do estrato } h \text{ possui o atributo, } A \subset U; \\ 0, & \text{caso contrário.} \end{cases}$$

# Parâmetros populacionais no estrato $h$

- O **total** populacional da variável  $y$  no estrato  $h$  é

$$T_{h,y} = \sum_{i \in U_h} y_i = N_{h,A},$$

onde  $N_{h,A}$  representa o **número de unidades populacionais** com o **atributo** de interesse.

- A **média populacional** no estrato  $h$  é dada por:

$$\bar{Y}_{h,y} = \frac{1}{N_h} \sum_{i \in U_h} y_i = \frac{T_h}{N_h} = \frac{N_{h,A}}{N_h} = P_h$$

- $P_h$  corresponde à **proporção**  $P$  de unidades da população que têm o atributo de interesse.
  - O parâmetro **proporção** é aqui representado pela letra  $P$  maiúscula, lembrando que  $P()$  maiúscula denota **probabilidade**.
- Uma *proporção* pode assumir valores variando entre 0, quando nenhuma unidade da população tem o atributo investigado, até 1, quando todas as unidades possuem esse atributo.

# Parâmetros populacionais no estrato $h$

- A **variância** populacional para  $y$  assumindo somente valores 0 ou 1 pode ser simplificada:

$$S_{h,y}^2 = \frac{1}{N_h - 1} \left( \sum_{i \in U_h} y_i^2 - N_h \bar{Y}_h^2 \right) = \frac{1}{N_h - 1} (N_h P_h - N_h P_h^2) = \frac{N_h}{N_h - 1} P_h (1 - P_h)$$

- A **variância** populacional de  $y$  pode também ser definida como

$$Var_{h,y} = P_h(1 - P_h).$$

- Tanto  $S_{h,y}^2$  como  $Var_{h,y}$  representam a dispersão da distribuição dos valores de  $y$  na população do estrato  $h$ .
- Para populações grandes ( $N_h \rightarrow \infty$ ), (é fácil verificar que)  
 $S_{h,y}^2 \doteq P_h(1 - P_h) = Var_y$ .

# Parâmetros populacionais globais

Na **AES** podemos escrever os parâmetros populacionais globais como abaixo.

- o **total populacional** de elementos com o atributo

$$T = \sum_{h=1}^H N_h P_h = N_A.$$

- A **proporção populacional**

$$P = \sum_{h=1}^H W_h P_h.$$

- A **variância populacional** sob **AASc** dentro dos estratos

$$Var_y = \sum_{i \in U} \frac{(y_i - \bar{Y})^2}{N} = \sum_{h=1}^H W_h Var_h + \sum_{h=1}^H W_h (P_h - P) = \sum_{h=1}^H W_h P_h (1 - P_h) + \sum_{h=1}^H W_h (P_h - P),$$

ou (lembrando)

$$S_y^2 = \frac{N}{N-1} Var_y.$$



# Estimadores de parâmetros globais

- Para o **total populacional** global de elementos com o atributo

$$\hat{T} = \sum_{h=1}^H N_h \hat{P}_h = N_A,$$

- onde  $\hat{P}_h = \frac{t_{h,y}}{n_h}$  é a proporção estimada no estrato  $h$ ,
- e  $t_{h,y} = \sum i \in s_h y_i$  o total de elementos com o atributo na amostra.
- A **proporção populacional**

$$P = \sum_{h=1}^H W_h \hat{P}_h.$$

- **Variância do estimador** da proporção nos estratos.
  - Sob **AASs** dentro,  $Var(\hat{P}_h) = \left( \frac{1}{n_h} - \frac{1}{N_h} \right) \frac{N_h}{N_h - 1} P_h (1 - P_h)$ .
  - Sob **AASc** dentro,  $Var(\hat{P}_h) = \frac{\hat{P}_h (1 - \hat{P}_h)}{n_h}$ .

# Variancia do estimador da proporção global

A **variância** do estimador da **proporção** populacional global é dada,

- sob **AASs** dentro dos estratos

$$Var_{AES_s}(\hat{P}_{AES}) = \sum_{h=1}^H W_h^2 \left( \frac{N_h - n_h}{N_h - 1} \right) \frac{P_h(1 - P_h)}{n_h};$$

- sob **AASc** dentro dos estratos

$$Var_{AES_c}(\hat{P}_{AES}) = \sum_{h=1}^H W_h^2 \frac{P_h(1 - P_h)}{n_h}.$$

# Estimador da variância do estimador da proporção

Usando os resultados acima podemos mostrar que o **estimador da variância** do estimador da proporção são dados abaixo.

- sob **AASs** dentro dos estratos

$$\widehat{Var}_{AES_s}(\widehat{P}_{AES}) = \sum_{h=1}^H W_h^2 \left( \frac{N_h - n_h}{N_h - 1} \right) \frac{\widehat{P}_h(1 - \widehat{P}_h)}{n_h - 1}.$$

- sob **AASc** dentro dos estratos

$$\widehat{Var}_{AES_c}(\widehat{P}_{AES}) = \sum_{h=1}^H W_h^2 \frac{\widehat{P}_h(1 - \widehat{P}_h)}{n_h - 1}.$$

## Exemplo 1

### Exercício 4.4 (Bolfarine e Bussab)

- 4.4** Planejou-se uma amostragem estratificada com reposição para estimar a porcentagem de famílias tendo conta em caderneta de poupança e também da quantidade investida. De uma pesquisa passada, têm-se estimativas para as proporções  $P_h$  e para os desvios padrões das quantidades investidas,  $\sigma_h$ , conforme descrito na tabela abaixo.

$h$	$W_h$	$P_h$	$\sigma_h$
1	0,6	0,20	9
2	0,3	0,40	18
3	0,1	0,60	52


Calcule os menores  $n$  e  $n_h$  que satisfaçam, com custo constante:

- A proporção populacional deve ser estimada com erro padrão igual a 0,02;
- A quantidade média investida deve ser estimada com erro padrão igual a R\$ 2,00.

Qual dos tamanhos, em (a) ou em (b), você usaria na pesquisa? Por quê?

## Exemplo 1

Exercício 4.4 (Bolfarine e Bussab)

a. Erro padrão (): No caso de  $\bar{y}$ ,

$$EP(\bar{y}) = \sqrt{Var(\bar{y})} = DP(\bar{y}).$$

- Como definir  $V$  em função de  $EP_{AES}(\hat{P}_{AES})$ ?
- Derivar uma expressão para  $n$  mínimo usando

$$V \geq Var_{AES}(\hat{P}_{AES}).$$

b. Como definir  $V$  em função de  $EP_{AES}(\bar{y}_{AES})$ ?

Interpretação e conclusão.

## Exemplo 2

### Exercício 4.20 (Bolfarine e Bussab)

**4.20** Uma cadeia de lojas está interessada em estimar, dentro das contas a receber, a proporção das que dificilmente serão recebidas. Para reduzir o custo da amostragem, usou-se AE com cada loja num estrato. Os dados obtidos foram os seguintes:

$h$	$N_h$	$n_h$	$\hat{P}_h$
1	60	15	0,30
2	40	10	0,20
3	100	20	0,40
4	30	6	0,10

onde  $N_h$  é o número de contas a receber,  $n_h$  é o tamanho da amostra e  $\hat{P}_h$  é a proporção de contas problemáticas. Dê uma estimativa para a proporção total de quatro lojas e um intervalo de confiança de 95% para a mesma.

## Exemplo 2

### Exercício 4.20 (Bolfarine e Bussab)

- Estimativa pontual  $\hat{P}_{AES}$ . Qual o valor? Interprete.
- Para o intervalo de confiança:
  - erro absoluto:

$$e = z_{\alpha/2} \times \sqrt{\hat{P}_{AES}}.$$

- Quais os limites?


$$IC\left(\hat{P}_{AES}; 1 - \alpha\right) = [\text{?}; \text{?}] .$$

- Interprete.

## Para casa

- Se preparar para a avaliação parcial da área.
- Continuar exercícios do livro 'Amostragem: Teoria e Prática Usando R'  
<https://amostragemcomr.github.io/livro/estrat.html#exerc11>
- Fazer exercícios da lista 1.
- Rever os slides.

## Próxima aula

- Amostragem Estratificada
  - Avaliação
  - Laboratório de 



# Muito obrigado!



Fonte: imagem do livro *Combined Survey Sampling Inference: Weighing of Basu's Elephants: Weighing Basu's Elephants*.

# Resumo da notação

## Estimação de Proporções na AES

Parâmetros	no estrato $h$
<b>total</b>	$T_{h,y} = \sum_{i \in U_h} y_i = N_{h,A}$
<b>proporção</b>	$\bar{Y}_{h,y} = \frac{1}{N_h} \sum_{i \in U_h} y_i = \frac{T_h}{N_h} = \frac{N_{h,A}}{N_h} = P_h$
<b>variância</b>	$Var_{h,y} = P_h(1 - P_h)$
<b>variância</b>	$S_{h,y}^2 = \frac{N_h}{N_h - 1} P_h(1 - P_h)$
Parâmetros	globais
<b>total</b>	$T = \sum_{h=1}^H N_h P_h = N_A$
<b>proporção</b>	$P = \sum_{h=1}^H W_h P_h$
<b>variância</b>	$Var_y = \sum_{h=1}^H W_h P_h(1 - P_h) + \sum_{h=1}^H W_h (P_h - P)^2$
<b>variância</b>	$S_y^2 = \frac{N}{N-1} Var_y$

# Resumo da notação

Estimadores	de parâmetros globais
total	$\hat{T} = \sum_{h=1}^H N_h \hat{P}_h = N_A$
proporção	$\hat{P} = \sum_{h=1}^H W_h \hat{P}_h$
Variância do estimador	da proporção nos estratos.
Sob AASs dentro	$Var(\hat{P}_h) = \left( \frac{1}{n_h} - \frac{1}{N_h} \right) \frac{N_h}{N_h - 1} P_h (1 - P_h)$
Sob AASc dentro	$Var(\hat{P}_h) = \frac{\hat{P}_h (1 - \hat{P}_h)}{n_h}$

# Resumo da notação

Variância	do estimador da proporção global
sob <b>AASs</b> dentro dos estratos	$Var_{AES_s}(\hat{P}_{AES}) = \sum_{h=1}^H W_h^2 \left( \frac{N_h - n_h}{N_h - 1} \right) \frac{P_h(1 - P_h)}{n_h}$
sob <b>AASc</b> dentro dos estratos	$Var_{AES_c}(\hat{P}_{AES}) = \sum_{h=1}^H W_h^2 \frac{P_h(1 - P_h)}{n_h}$
Estimador da variância	do estimador da proporção global
sob <b>AASs</b> dentro dos estratos	$\hat{V}ar_{AES_s}(\hat{P}_{AES}) = \sum_{h=1}^H W_h^2 \left( \frac{N_h - n_h}{N_h - 1} \right) \frac{\hat{P}_h(1 - \hat{P}_h)}{n_h - 1}$
sob <b>AASc</b> dentro dos estratos	$\hat{V}ar_{AES_c}(\hat{P}_{AES}) = \sum_{h=1}^H W_h^2 \frac{\hat{P}_h(1 - \hat{P}_h)}{n_h - 1}$

# Resumo da notação

- Tamanho mínimo de amostra para **estimação da proporção** populacional

Alocação	AASc dentro dos estratos
$AES_{un}$	$n \geq \frac{H \sum_{h=1}^H W_h^2 P_h (1-P_h)}{V}$
$AES_{pr}$	$n \geq \frac{\sum_{h=1}^H W_h P_h (1-P_h)}{V}$
$AES_{ne}$	$n \geq \frac{\left( \sum_{h=1}^H W_h \sqrt{P_h (1-P_h)} \right)^2}{V}$
$AES_{ot}$	$n \geq \frac{\left( \sum_{h=1}^H W_h \sqrt{P_h (1-P_h)} \sqrt{C_h} \right) \left( \sum_{h=1}^H W_h \sqrt{P_h (1-P_h)} / \sqrt{C_h} \right)}{V}$

# Resumo da notação

- Tamanho mínimo de amostra para **estimação da proporção** populacional

Alocação	AASs dentro dos estratos
$AES_{un}$	$n \geq \frac{H \sum_{h=1}^H W_h^2 P_h (1-P_h)}{V + \frac{1}{N} \sum_{h=1}^H W_h P_h (1-P_h)}$
$AES_{pr}$	$n \geq \frac{\sum_{h=1}^H W_h P_h (1-P_h)}{V + \frac{1}{N} \sum_{h=1}^H W_h P_h (1-P_h)}$
$AES_{ne}$	$n \geq \frac{\left( \sum_{h=1}^H W_h \sqrt{P_h (1-P_h)} \right)^2}{V + \frac{1}{N} \sum_{h=1}^H W_h P_h (1-P_h)}$
$AES_{ot}$	$n \geq \frac{\left( \sum_{h=1}^H W_h \sqrt{P_h (1-P_h)} \sqrt{\mathcal{C}_h} \right) \left( \sum_{h=1}^H W_h \sqrt{P_h (1-P_h)} / \sqrt{\mathcal{C}_h} \right)}{V + \frac{1}{N} \sum_{h=1}^H W_h P_h (1-P_h)}$

# Referências

- Amostragem: Teoria e Prática Usando o R

\* Cochran(1977)

# Revisão Estimação de Proporções na AAS



# Estimação de Proporções

Um caso especial do que já vimos ocorre quando a variável  $y$  indicadora de uma unidade populacional  $i$  tem ou não uma determinada **característica** ou **atributo de interesse**, ou pertence a um determinado grupo especificado de unidades da população:

- Migrantes entre os habitantes de determinada região.
- Estabelecimentos agropecuários de produção leiteira numa localidade.
- Estudantes do sexo feminino em escolas.
- Sondagens eleitorais, parcela dos eleitores pretende votar em determinado candidato.

Definimos a variável  $y$ , para cada unidade  $i$  da população, assumindo um de dois valores possíveis:

$$y_i = I(i \in A) = \begin{cases} 1, & \text{se a unidade } i \text{ possui o atributo, } A \subset U; \\ 0, & \text{caso contrário.} \end{cases}$$

# Parâmetros populacionais

- O **total** populacional da variável  $y$  é

$$T = \sum_{i \in U} y_i = N_A,$$

onde  $N_A$  representa o **número de unidades populacionais** com o **atributo** de interesse.

- Variáveis indicadoras são usadas quando se quer tabular frequências de respostas a uma pergunta categórica.
- Para respostas sendo valores inteiros de 1 a  $H$ , onde  $H$  representa o número de categorias de resposta da pergunta.
  - Ex.: na pergunta 'Qual é o sexo do morador', há duas categorias de resposta ( $H = 2$ ): 1 (=Feminino) e 2 (=Masculino).
  - Para contar o número de pessoas por sexo na população, definimos:  $y_{1i} = I[Sexo(i) = 1]$  e  $y_{2i} = I[Sexo(i) = 2]$ .

# Parâmetros populacionais

Sabemos que quando a variável  $y$  é do tipo indicadora, a **média populacional** dada por:

$$\bar{Y} = \frac{1}{N} \sum_{i \in U} y_i = \frac{T}{N} = \frac{N_A}{N} = P$$

- $P$  corresponde à **proporção**  $P$  de unidades da população que têm o atributo de interesse.
  - O parâmetro **proporção** é aqui representado pela letra  $P$  maiúscula, lembrando que  $P()$  maiúscula denota **probabilidade**.
- Uma *proporção* pode assumir valores variando entre 0, quando nenhuma unidade da população tem o atributo investigado, até 1, quando todas as unidades possuem esse atributo.
  - Muitas vezes é interessante expressar a **proporção** sob forma de **porcentagem** podendo então variar de 0% até 100%.

# Parâmetros populacionais

- A **variância** populacional para  $y$  assumindo somente valores 0 ou 1 pode ser simplificada:

$$S_y^2 = \frac{1}{N-1} \left( \sum_{i \in U} y_i^2 - N\bar{Y}^2 \right) = \frac{1}{N-1} (NP - NP^2) = \frac{N}{N-1} P(1-P)$$

- A **variância** populacional de  $y$  pode também ser definida como  $Var_y = P(1-P)$ .
  - Tanto  $S_y^2$  como  $Var_y$  representam a dispersão da distribuição dos valores de  $y$  na população.
  - Para populações grandes ( $N \rightarrow \infty$ ), (é fácil verificar que)  $S_y^2 \doteq P(1-P) = Var_y$ .

- O **Coefficiente de Variação** ( $CV$ ),

$$CV_y = \frac{\sqrt{var_y}}{\bar{Y}} = \sqrt{P(1-P)/P^2} = \sqrt{(1-P)/P}.$$

# Estimação sob amostragem aleatória simples AAS

- Seja  $s$  uma amostra observada sob **AAS** de tamanho  $n$  de uma população com  $N$  unidades, onde se observa uma variável indicadora  $y$ .
  - Podemos obter estimadores para os parâmetros populacionais de  $y$  adaptando os estimadores gerais de total e média apresentados no capítulo anterior.
- O **total de unidades da amostra** com o **atributo** de interesse,  $n_A$ , será dado pela soma amostral:

$$t_y = \sum_{i \in s} y_i = n_A$$

- O **total estimado** de unidades na população com o **atributo** de interesse,  $N_A$ , é estimado usando:

$$\hat{T}_{AAS} = N \times t_y/n = N \times n_A/n = \hat{N}_A$$

- Como sabemos este estimador é não viciado sob AAS para qualquer variável  $y$ , logo é **ENV** também quando  $y$  é do tipo indicadora.

# Estimação sob amostragem aleatória simples AAS

- A **proporção amostral** de unidades que possuem o atributo de interesse é dada pela **média amostral**:

$$\bar{y} = \frac{1}{n} \sum_{i \in s} y_i = \frac{n_A}{n} = \hat{P}$$

- Pode-se facilmente verificar que  $\hat{p}$  é um *estimador não viciado* para a *proporção* populacional  $p$ , pois:

$$E_{AAS}(\hat{P}) = E_{AAS}(\bar{y}) = \bar{Y} = P$$

- Mostrar (?)

# Estimação sob amostragem aleatória simples

## COM REPOSIÇÃO - AASc

- sob amostragem aleatória simples com reposição \*\*AAS

A variância da proporção amostral sob AASC é dada por:

$$Var_{AASC}(\hat{P}) = \frac{Var_y}{n} = \frac{P(1 - P)}{n}$$

ou

$$\hat{S}_y^2 = \frac{n}{n - 1} \hat{P}(1 - \hat{P})$$

- Sob AASc,  $\hat{S}_y^2$  é um ENV para  $Var_y$ . Assim

$$\widehat{Var}_{AASc}(\hat{P}) = \frac{\hat{P}(1 - \hat{P})}{n - 1}$$

# Estimação sob amostragem aleatória simples

## COM REPOSIÇÃO - AASc

- O **estimador do total** de unidades na população que possuem o **atributo** de interesse,  $N_A = NP$  é dado por

$$\widehat{N}_A = N\widehat{P}.$$

- A **variância do estimador** de  $N_A$ ,

$$V_{AASc}(\widehat{N}_A) = N^2 \frac{P(1 - P)}{n}$$

pode ser estimada por

$$\widehat{V}ar_{AASc}(\widehat{N}_A) = N^2 \frac{\widehat{P}(1 - \widehat{P})}{n - 1}.$$



# Estimação sob amostragem aleatória simples

## SEM REPOSIÇÃO AASS

- No caso de  $s$  ser obtida por seleção do tipo **AASS**,
  - a **soma amostral**  $t_y$ , a **proporção amostral**  $\hat{p}$  e a **variância amostral**  $\hat{S}_y^2$  têm a **mesma forma** que na **AASC**.
  - Os estimadores para o total populacional  $N_A$  e a proporção populacional  $P$  são os mesmos.
- A **diferença** é que na **AASS** a variância amostral  $\hat{S}_y^2$  é um **ENV** para  $S_y^2$  e não para  $Var_y$ .
  - Também são **diferentes** as expressões para as variâncias dos estimadores amostrais e seus correspondentes estimadores não viciados.

# Estimação sob amostragem aleatória simples

## SEM REPOSIÇÃO AASs

- Sabemos que as **variâncias** dos **estimadores** do **total** e da **média** são

$$Var_{AASs}(\hat{Y}) = N^2 \left( \frac{1}{n} - \frac{1}{N} \right) S_y^2 \quad \text{e} \quad Var_{AASs}(\bar{y}) = \left( \frac{1}{n} - \frac{1}{N} \right) S_y^2.$$

- No caso de variáveis  $y$  do tipo indicadoras, tem-se que as **variâncias** do **estimador** do **total** e da **proporção populacionais** são dadas por:

$$Var_{AASs}(\hat{N}_A) = N^2 \left( \frac{1}{n} - \frac{1}{N} \right) \frac{N}{N-1} P(1-P) \quad \text{e} \quad Var_{AASs}(\hat{P}) = \left( \frac{1}{n} - \frac{1}{N} \right) \frac{N}{N-1} P(1-P).$$

- Se o número de unidades  $N$  é suficientemente grande, tem-se que  $Var_{AAS}(\hat{p}) \doteq \frac{P(1-P)}{n}$ , resultando em desempenhos similares entre **AASs** e **AASC** na estimação da proporção populacional.
  - Intuitivamente, isso ocorre porque a probabilidade de seleção repetida sob **AASc** tende a ser muito pequena no caso de populações muito grandes.

# Estimação sob amostragem aleatória simples

## SEM REPOSIÇÃO AASs

Utilizando  $\hat{S}_y^2$  como estimador não viciado para  $S_y^2$  chega-se aos estimadores para as variâncias dos estimadores de total e proporção:

$$\hat{V}ar_{AAS}(\hat{N}_A) = N^2 \left( \frac{1}{n} - \frac{1}{N} \right) \frac{n\hat{P}(1 - \hat{P})}{n - 1}$$

e

$$\hat{V}ar_{AAS}(\hat{P}) = \left( \frac{1}{n} - \frac{1}{N} \right) \frac{n\hat{P}(1 - \hat{P})}{n - 1}$$

# Distribuição amostral EXATA de estimadores

## COM REPOSIÇÃO AASc

- Na **AASc**, a **soma amostral**  $t_y = n_A \sim \text{Bernoulli}(P)$ , portanto,  
 $t_y = n_A \sim \text{Binomial}(n, P)$ ,

$$E_{AASc}(n_A) = np \quad \text{e} \quad V_{AASc}(n_A) = nP(1 - P)$$

- Da mesma forma o **valor esperado** e a **variância** de  $\hat{P}$ :

$$E_{AASc}(\hat{P}) = E_{AASc}\left(\frac{n_A}{n}\right) = P \quad \text{e} \quad \text{Var}_{AASc}(\hat{P}) = \frac{P(1 - P)}{n}$$

Outro resultado importante da **distribuição** de probabilidades **exata** de  $\hat{P}$ ,

$$P\left(\hat{P} = \frac{v}{n}\right) = P(n_A = v) = \binom{n}{v} P^v (1 - P)^{n-v}, \quad \forall v = 0, 1, 2, \dots, n.$$

Esta distribuição corresponde apenas a uma transformação escalar da distribuição  $\text{Binomial}(n, p)$ , onde a contagem de sucessos ( $n_A$ ) é dividida pelo número de sorteios ( $n$ ).

# Distribuição amostral EXATA de estimadores

## SEM REPOSIÇÃO AASs

- Sob **AASs**, temos  $n_A \sim \text{Hipergeométrica}(N, N_A, n)$ ,  $n$  sorteios são feitos da população sem reposição.
  - O **número total de amostras** aleatórias simples sem reposição de tamanho  $n$ ,  $\binom{N}{n}$ ;
  - o número dessas amostras com exatamente  $v$  unidades com a característica em estudo e  $n - v$  unidades sem essa característica pode ser calculado por  $\binom{N_A}{v} \binom{N-N_A}{n-v}$ , assim a **distribuição** de probabilidades de  $t_y = n_A$  é dada por:

$$P(n_A = v) = \frac{\binom{N_A}{v} \binom{N-N_A}{n-v}}{\binom{N}{n}}, \quad \forall v = 0, 1, 2, \dots, \min(n; N_A)$$

# Distribuição amostral EXATA de estimadores

## SEM REPOSIÇÃO AASs

- Assim fica também determinada a distribuição exata de probabilidades do estimador  $\hat{p}$ , que é a mesma de  $n_A$  com os valores possíveis divididos pelo tamanho da amostra  $n$ .

Consequentemente tem-se que o valor esperado de unidades com o atributo de interesse na amostra e sua variância serão dados por:

$$E_{AAS}(n_A) = n \frac{n_A}{N} = nP \quad \text{e} \quad Var_{AAS}(n_A) = nP(1 - P) \frac{N - n}{N - 1}$$

Para o estimador,  $\hat{p} = n_A/n$ , da proporção de unidades com o atributo de interesse na população tem-se:

$$E_{AAS}(\hat{p}) = p \quad \text{e} \quad V_{AAS}(\hat{p}) = \left( \frac{1}{n} - \frac{1}{N} \right) S_y^2$$

# Intervalos de confiança para proporções na amostragem aleatória simples

- Na AAS, tanto com ou sem reposição, são conhecidas as distribuições exatas para o estimador  $\hat{p}$ .
  - Portanto, é possível obter os limites inferior e superior para intervalos de confiança para a proporção  $p$ , com um nível de significância  $\alpha$  fixado.

Para isso, no caso de **AASc**, é necessário resolver o sistema de equações determinando os valores de  $\hat{p}_I$  e  $\hat{p}_S$  que satisfaçam:

$$\begin{cases} \sum_{v=0}^{n_A} \binom{n}{v} \hat{p}_S^v (1 - \hat{p}_S)^{n-v} = \alpha/2 \\ \sum_{v=n_A}^n \binom{n}{v} \hat{p}_I^v (1 - \hat{p}_I)^{n-v} = \alpha/2 \end{cases}$$

No caso da **AASs**, o sistema a ser resolvido é baseado na distribuição Hipergeométrica como se segue:

$$\begin{cases} \sum_{v=0}^{n_A} \frac{\binom{N\hat{p}_S}{v} \binom{N-N\hat{p}_S}{n-v}}{\binom{N}{n}} = \alpha/2 \\ \sum_{v=n_A}^n \frac{\binom{N\hat{p}_I}{v} \binom{N-N\hat{p}_I}{n-v}}{\binom{N}{n}} = \alpha/2 \end{cases}$$

Em ambos os casos  $1 - \alpha$  é o *nível de confiança* desejado. Por exemplo, para intervalos de 95% de confiança, deve-se usar  $\alpha = 0,05$ .

# Intervalos de confiança para proporções na amostragem aleatória simples

- A solução desses sistemas costumava ser trabalhosa, exigindo aplicação de métodos iterativos que consumiam quantidade razoavelmente grande de recursos computacionais.
  - Atualmente, com o avanço dos métodos computacionais, esse problema pode facilmente ser resolvido, por exemplo, com o uso do R. Uma maneira é utilizar as funções *qbinom* e *qhyper* que podem calcular os quantis das distribuições Binomial e Hipergeométrica para  $\alpha/2$  e  $1 - \alpha/2$ .
- Além disso, há outros programas prontos facilmente utilizáveis como, por exemplo, as funções *binconf* e *confCI* incluídas, respectivamente, nos pacotes *Hmisc* e *prevalence* do R.
  - Essas funções estimam intervalos de confiança para vários métodos além do mostrado acima, como o da aproximação Normal, apresentado na próxima seção, além de outras abordagens.
- Há, também, no pacote *survey* uma função específica, *svyciprop*, para calcular intervalos de confiança para proporções. Uma característica interessante do pacote *survey* é que é possível determinar a utilização do



# Intervalos de confiança utilizando a aproximação Normal

- A distribuição do estimador da proporção,  $\hat{P}$ , pode ser aproximada pela distribuição Normal de probabilidade.
  - Pode ser utilizada mesmo no caso da **AAS** onde os  $y_i$  observados na amostra não são independentes, desde que se tenha valores de  $N$  e  $n$  suficientemente grandes e valor da fração amostral,  $f = \frac{n}{N}$ , pequeno.
- Sob estas condições pode-se considerar que:

$$\frac{\hat{p} - p}{\sqrt{V_{p(s)}(\hat{p})}} \approx N(0; 1)$$

# Intervalos de confiança utilizando a aproximação Normal

Os histogramas abaixo mostram os valores estimados da proporção  $P$  de unidades com uma determinada característica de interesse, a partir de 1.000 amostras aleatórias simples de tamanho  $n = 100$ , de uma população de tamanho  $N = 5.000$ , onde exatamente metade das unidades tem a característica de interesse ( $p = 1/2$ ). Valores normalizados de acordo com a aproximação acima.

Cochran(1977) mostra uma tabela, com alguns valores mínimos do total de unidades observadas na amostra,  $n_A$ , onde a aproximação Normal pode ser utilizada.

$p$	$n_A$	$n$
0,50	15	30
0,40	20	50
0,30	24	80
0,20	40	200
0,10	60	600
0,05	70	1.400
$\doteq 0$	80	$\infty$

A Tabela \@ref(tab:tabprop3) foi construída considerando um nível de significância de  $\alpha = 0,05$ , que é um valor comumente utilizado em muitas situações práticas. Tem-se, a partir daí, critérios práticos para assumir a utilização da aproximação Normal, notando-se que o tamanho mínimo da amostra requerido é de  $n = 30$ .

Nas condições estabelecidas para a validade da aproximação Normal, tem-se que  $S_y^2 \doteq \sigma_y^2 = p(1 - p)$ , portanto,  $V_{AAS}(\hat{p}) \doteq V_{AASC}(\hat{p})$ . Então, para os dois tipos de seleção, pode-se considerar o intervalo de confiança para a proporção como:

$$IC(p; 1 - \alpha) = \left[ \hat{p} - z_{\alpha/2} \sqrt{\frac{p(1 - p)}{n}} ; \hat{p} + z_{\alpha/2} \sqrt{\frac{p(1 - p)}{n}} \right]$$

Caso se deseje considerar o fator de correção para populações finitas, quando a fração amostral não possa ser considerada pequena e a seleção for sem reposição, a expressão do intervalo de confiança passa a ser:

$$IC(p; 1 - \alpha) = \left[ \hat{p} \mp z_{\alpha/2} \sqrt{\left( \frac{N - n}{N - 1} \right) \frac{p(1 - p)}{n}} \right]$$

Em Cochran(1977) também é apresentada uma *correção de continuidade* acrescentando a fração  $1/2n$  à margem de erro do intervalo de confiança pelo fato de se fazer uma aproximação de uma distribuição discreta (Binomial ou Hipergeométrica) pela distribuição Normal, que é contínua. Desse modo a expressão do intervalo de confiança passa a ser:

$$IC(p; 1 - \alpha) = \left[ \hat{p} \mp \left( z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} + \frac{1}{2n} \right) \right]$$

Ou considerando a correção para população finita:

$$IC(p; 1 - \alpha) = \left[ \hat{p} \mp \left( z_{\alpha/2} \sqrt{\left( \frac{N-n}{N-1} \right) \frac{p(1-p)}{n}} + \frac{1}{2n} \right) \right]$$

Nas aplicações práticas o valor da variância do estimador da proporção  $p$ , geralmente, não é conhecido. Assim o que se pode fazer é estimar um intervalo de confiança, substituindo  $S_y^2$  por  $\hat{S}_y^2$  na expressões anteriores:

$$\widehat{IC}(p; 1 - \alpha) = \left[ \hat{p} \mp \left( z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n-1}} + \frac{1}{2n} \right) \right]$$

$$\widehat{IC}(p; 1 - \alpha) = \left[ \hat{p} \mp \left( z_{\alpha/2} \sqrt{\left( \frac{N-n}{N} \right) \frac{\hat{p}\hat{q}}{n-1}} + \frac{1}{2n} \right) \right]$$

Note que o efeito da correção de continuidade tende rapidamente a ser nulo quando o tamanho da amostra,  $n$ , cresce. Para uma amostra de tamanho  $n = 50$  esse fator já é de apenas 1%, o que pode ser desprezível dependendo da proporção que estiver sendo estimada, porém é preciso muito cuidado pois quando se está trabalhando com proporções são valores, às vezes, bastante pequenos.

# Cálculo do tamanho da amostra

O tamanho de uma amostra aleatória simples a ser selecionada, como já foi visto no capítulo anterior, é calculado a partir da definição do erro amostral ou margem de erro admissível para o caso, do nível de confiança desejado e se a seleção for com ou sem reposição.

No caso de seleção com reposição, considerando uma margem de erro máxima admissível  $D$  com um nível de confiança  $1 - \alpha$ , basta utilizar a expressão da margem de erro:

$$D \leq z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \implies n \geq \frac{z_{\alpha/2}^2 p(1-p)}{D^2}$$

Para a seleção sem reposição, o tamanho da amostra é calculado como:

$$D \leq z_{\alpha/2} \sqrt{\left(\frac{N-n}{N-1}\right) \frac{p(1-p)}{n}} \implies n \geq \frac{z_{\alpha/2}^2 p(1-p)}{D^2 \frac{N-1}{N} + \frac{1}{N} z_{\alpha/2}^2 p(1-p)}$$

Considerando  $\frac{N-1}{N} \doteq 1$ , tem-se que:

$$D \leq z_{\alpha/2} \sqrt{\left(\frac{N-n}{N-1}\right) \frac{p(1-p)}{n}} \implies n \geq \frac{Np(1-p)}{ND^2/z_{\alpha/2}^2 + p(1-p)}$$

Uma maneira prática de calcular o tamanho da amostra para uma AAS em dois passos é calcular primeiro:

$$n_0 = \frac{z_{\alpha/2}^2 p(1-p)}{D^2}$$

E depois fazer:

$$n \geq \frac{n_0}{1 + n_0/N}$$

Note que  $n_0$  é equivalente ao tamanho da amostra para uma AASC e o valor de  $n$  para a AAS é obtido pela correção para população finita do valor  $n_0$ . Também pode-se concluir que quando o tamanho da população,  $N$ , é grande o fator  $n_0/N$  tende a se anular fazendo com que  $n \doteq n_0$ , ou seja, quando o tamanho da população é grande as amostras aleatórias simples com ou sem reposição são equivalentes.