

Análise bayesiana de pesquisas RDS com incerteza no desfecho

Lucas M. Moschen
Orientador: Luiz Max F. de Carvalho

Escola de Matemática Aplicada
Fundação Getulio Vargas

24 de junho de 2021

- ① Introdução
- ② Justificativa
- ③ Objetivos
- ④ Metodologia
- ⑤ Resultados preliminares
- ⑥ Cronograma

- ① Introdução
- ② Justificativa
- ③ Objetivos
- ④ Metodologia
- ⑤ Resultados preliminares
- ⑥ Cronograma

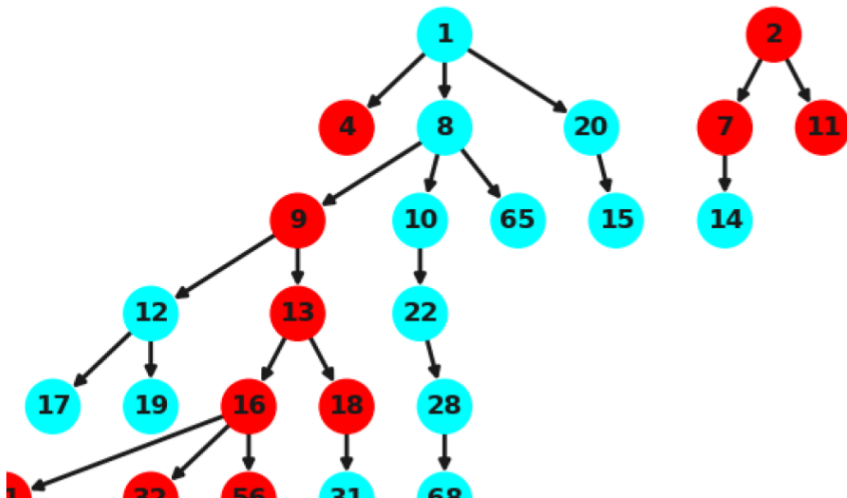
Populações de difícil acesso ou “escondidas”

- ▶ Não existe um esquema de amostragem clássico: tamanho e fronteiras da população são desconhecidos;
- ▶ Preocupações com a privacidade e medo de exposição: comportamento estigmatizado ou ilegal;
- ▶ Exemplos: Usuários de drogas pesadas, profissionais do sexo, pessoas em situação de rua, entre outros;
- ▶ Abordagens existentes de amostragem têm vários espaços para desenvolvimento.

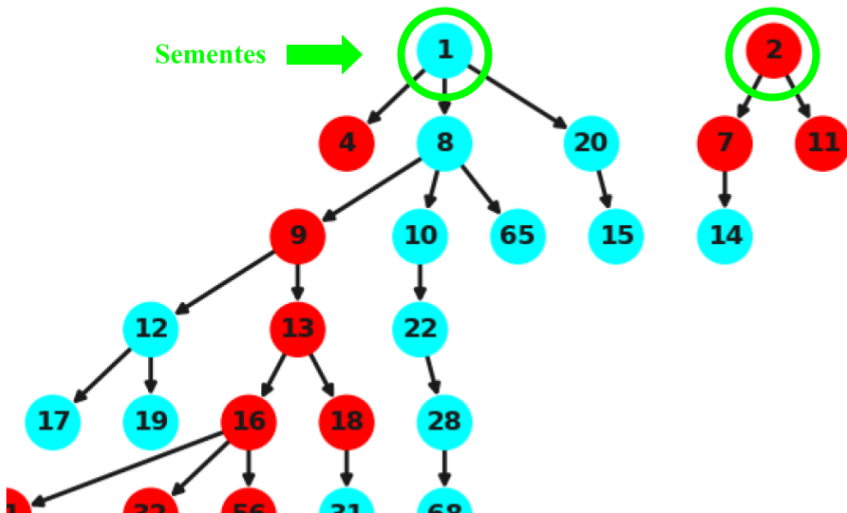
Respondent-driven sampling

- ▶ Proposta em [[Heckathorn, 1997](#)] como uma abordagem de estimar proporções em uma população de difícil acesso;
- ▶ Teoria baseada em cadeias de Markov;
- ▶ [[Crawford, 2016](#)] modela como uma rede com interações e define uma distribuição de probabilidade sobre o subgrafo observado;
- ▶ Amostragem sem reposição.

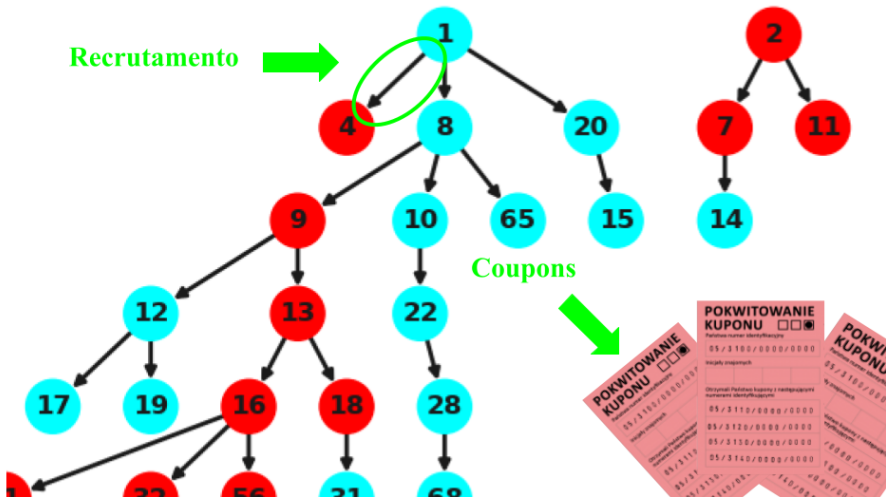
Estudo RDS: refugiados e ativistas na Síria



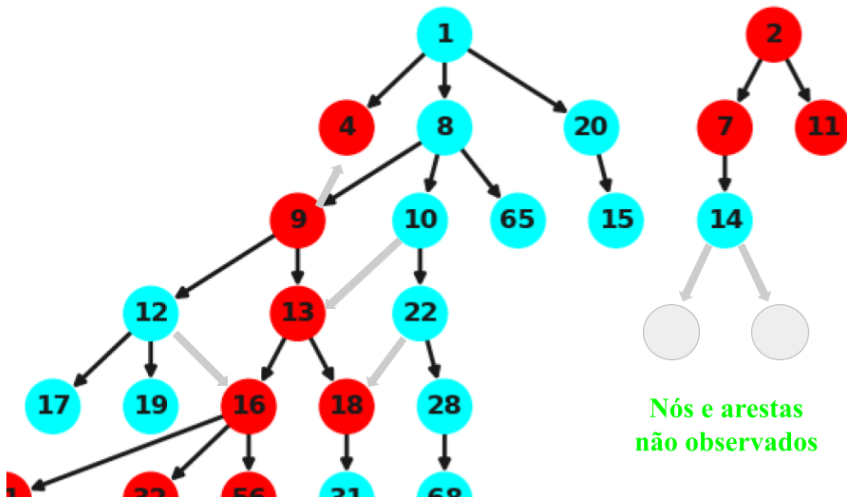
Estudo RDS: refugiados e ativistas na Síria



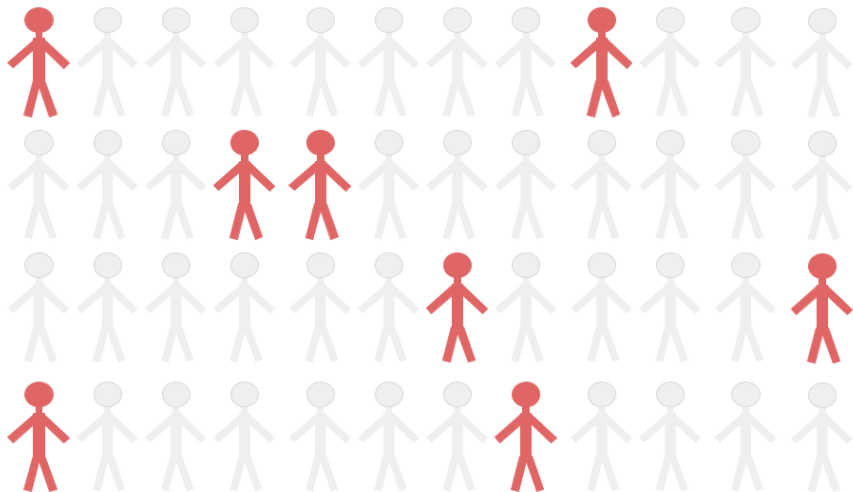
Estudo RDS: refugiados e ativistas na Síria



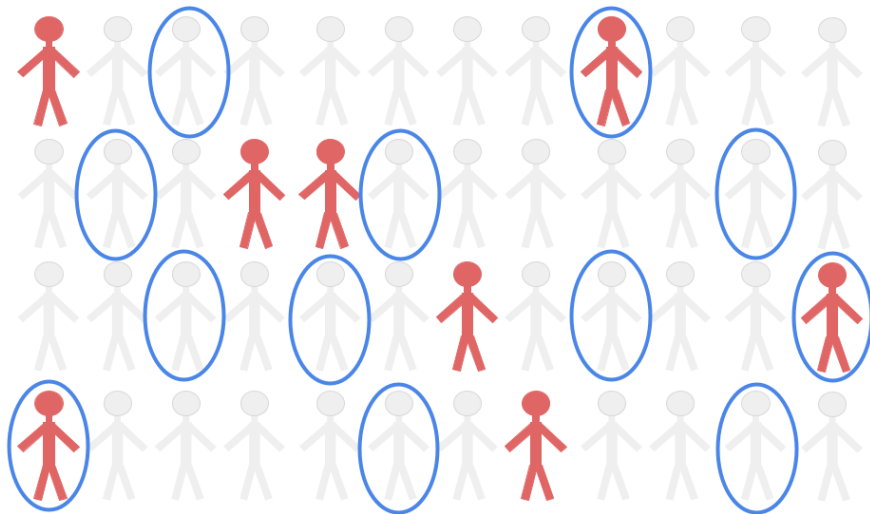
Estudo RDS: refugiados e ativistas na Síria



Estimação de prevalência com testes imperfeitos



Estimação de prevalência com testes imperfeitos



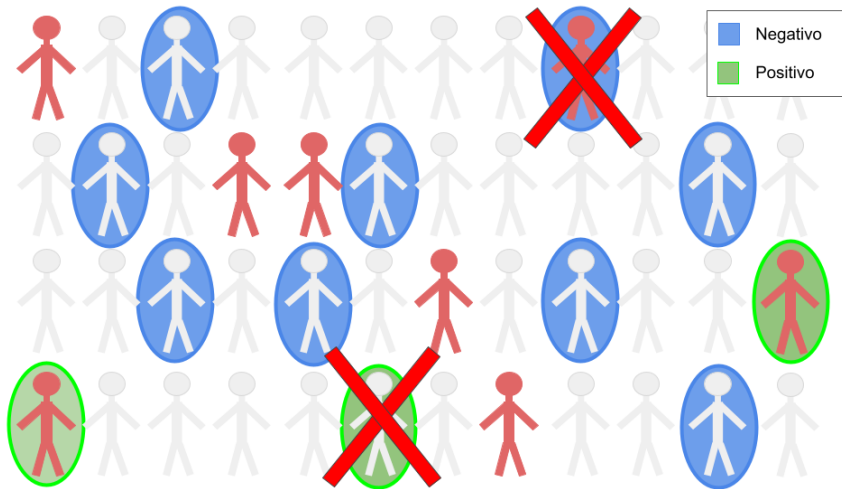
Estimação de prevalência com testes imperfeitos



Especificidade e Sensibilidade

Especificidade: Probabilidade de resultado negativo nos não doentes.

Sensibilidade: Probabilidade de resultado positivo nos doentes.



- ▶ Interpretação baseada no grau de crença em uma afirmação por um indivíduo;
- ▶ A fórmula de Bayes relaciona a probabilidade de um parâmetro após observar novos dados com a evidência e a informação prévia sobre ele;
- ▶ Permite a quantificação da incerteza de uma forma direta, dado que o processo não precisa ser aleatório.

- 1 Introdução
- 2 Justificativa
- 3 Objetivos
- 4 Metodologia
- 5 Resultados preliminares
- 6 Cronograma

- ▶ Populações escondidas são sub-representadas em pesquisas nacionais e têm maior risco de abuso de drogas ou contrair infecções sexualmente transmissíveis;
- ▶ O tópico tem vários gaps na estatística e abordagens de regressão para estimar prevalência considerando a estrutura de rede podem ser construídas [[Bastos et al., 2012](#)].

- 1 Introdução
- 2 Justificativa
- 3 Objetivos**
- 4 Metodologia
- 5 Resultados preliminares
- 6 Cronograma

Este trabalho tem o objetivo de estudar o problema da estimação de prevalência em uma estrutura de rede RDS considerando a Especificidade e a Sensibilidade (desconhecidas) do diagnóstico.

Também se pretende aplicar esse framework de forma eficiente, comparando algoritmos Monte Carlo e Aproximações de Laplace.

- 1 Revisão bibliográfica, descrição matemática do problema e propagação da incerteza com métodos Bayesianos;
- 2 Distribuição conjunta a priori da especificidade e sensibilidade;
- 3 Implementação eficiente utilizando pacotes estatísticos, como *rstanarm* e *INLA*;
- 4 Análise de estudos epidemiológicos RDS.

- 1 Introdução
- 2 Justificativa
- 3 Objetivos
- 4 Metodologia
- 5 Resultados preliminares
- 6 Cronograma

Pesquisa bibliográfica

A fundamentação teórica se dará por meio de artigos nos tópicos indicados na introdução: RDS, estimativa de prevalência por meio de regressão, e estatística bayesiana.

Recursos técnicos

Toda a programação necessária será feita nas linguagens de programação *Python* e *R*.

Estudo formal

Disciplinas do Doutorado em Modelagem Matemática da EMap: Estatística Bayesiana e Ciências de Redes.

- 1 Introdução
- 2 Justificativa
- 3 Objetivos
- 4 Metodologia
- 5 Resultados preliminares
- 6 Cronograma

Um resultado preliminar

Considere o modelo [Gelman and Carpenter, 2020]:

$$y \sim \text{Binomial}(n, p),$$
$$p = \theta \gamma_s + (1 - \theta)(1 - \gamma_e).$$

Informação a priori sobre especificidade e sensibilidade:

$$y_{negativos} \sim \text{Binomial}(n_{\gamma_e}, \gamma_e),$$
$$y_{positivos} \sim \text{Binomial}(n_{\gamma_s}, \gamma_s).$$

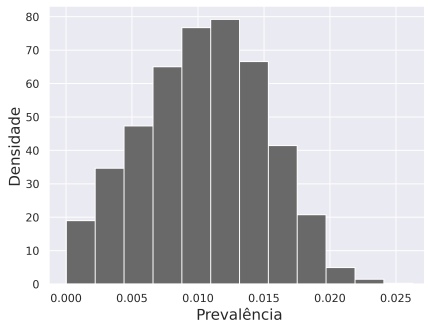
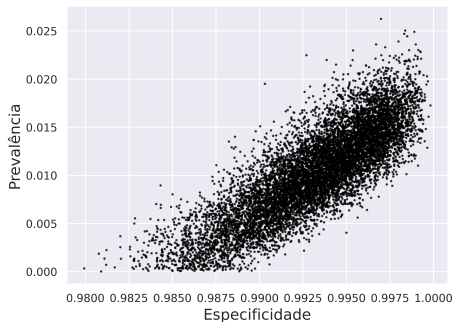
Distribuições a priori para os parâmetros:

$$\gamma_e \sim \text{Beta}(a_e, b_e),$$
$$\gamma_s \sim \text{Beta}(a_s, b_s),$$
$$\theta \sim \text{Beta}(a_\theta, b_\theta).$$

Um resultado preliminar

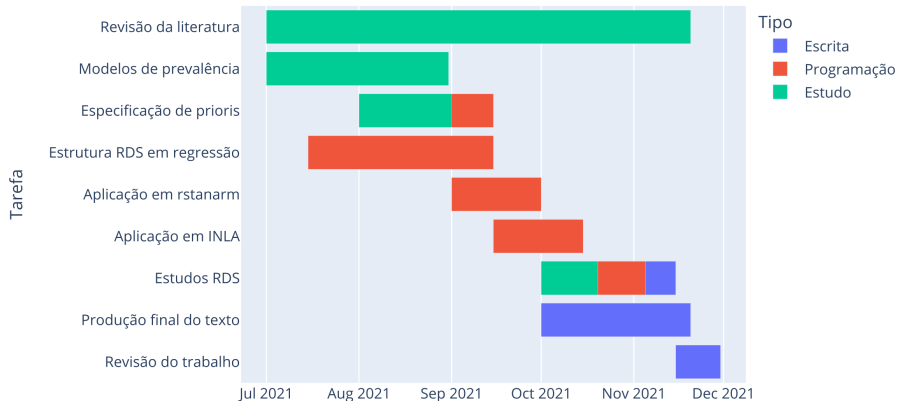
Dados de soroprevalência de COVID:

$$\begin{aligned}y/n &= 50/3330, \\ y_{\text{negativos}}/n_{\gamma_e} &= 399/401, \\ y_{\text{positivos}}/n_{\gamma_s} &= 103/122.\end{aligned}$$



- 1 Introdução
- 2 Justificativa
- 3 Objetivos
- 4 Metodologia
- 5 Resultados preliminares
- 6 Cronograma

Cronograma





Albuquerque, E. M. d. et al. (2009).

Avaliação da técnica de amostragem respondent-driven sampling na estimação de prevalências de doenças transmissíveis em populações organizadas em redes complexas.

PhD thesis.



Bastos, L. S., Pinho, A. A., Codeço, C., and Bastos, F. I. (2012).

Binary regression analysis with network structure of respondent-driven sampling data.



Crawford, F. W. (2016).

The graphical structure of respondent-driven sampling.

Sociological Methodology, 46(1):187–211.

References II



Crawford, F. W., Wu, J., and Heimer, R. (2018).

Hidden population size estimation from respondent-driven sampling: a network approach.

Journal of the American Statistical Association, 113(522):755–766.



Deaux, E. and Callaghan, J. W. (1985).

Key informant versus self-report estimates of health-risk behavior.

Evaluation Review, 9(3):365–368.



Gelman, A. and Carpenter, B. (2020).

Bayesian analysis of tests with unknown specificity and sensitivity.

Journal of the Royal Statistical Society: Series C (Applied Statistics), 69(5):1269–1283.

References III



Goodman, L. A. (1961).

Snowball Sampling.

The Annals of Mathematical Statistics, 32(1):148–170.



Heckathorn, D. D. (1997).

Respondent-driven sampling: A new approach to the study of hidden populations.

Social Problems, 44(2):174–199.






Heckathorn, D. D. (2002).

Respondent-driven sampling ii: deriving valid population estimates from chain-referral samples of hidden populations.

Social problems, 49(1):11–34.

References IV

-  McInturff, P., Johnson, W. O., Cowling, D., and Gardner, I. A. (2004).
Modelling risk when binary outcomes are subject to error.
Statistics in medicine, 23(7):1095–1109.
-  Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M., and Zwinderman, A. H. (2005).
Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews.
Journal of clinical epidemiology, 58(10):982–990.
-  Robert, C. (2007).
The Bayesian choice: from decision-theoretic foundations to computational implementation.
Springer Science & Business Media.

References V

 Rue, H., Martino, S., and Chopin, N. (2009).

Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations.

Journal of the royal statistical society: Series b (statistical methodology), 71(2):319–392.

 Stan Development Team (2021).

Rstanarm developer notes.

Rstanarm website.

 Statisticat, L. (2016).

Laplacesdemon: A complete environment for bayesian inference within r.

R Package version, 17:2016.

References VI



Volz, E. and Heckathorn, D. D. (2008).

Probability based estimation theory for respondent driven sampling.

Journal of official statistics, 24(1):79.



Watters, J. K. and Biernacki, P. (1989).

Targeted sampling: Options for the study of hidden populations.

Social Problems, 36(4):416–430.

Muito obrigado!