有意抽出でも N 増えれば誤差減少

誤差は二項分布推定のほうが小さいが、信頼区間は誤差相関モデルのほうが正確

スノーボール・サンプルによる比率の推定

太郎丸 博[1]

@tarohmaru

taroumaru.hiroshi.7u@kyoto-u.ac.jp

[1] 京都大学 文学研究科

1 問題

- 有意抽出したデータから一般化したい (太郎丸 2021)
- スノーボールサンプリングした場合の誤差の大きさは?
- 誤差が母集団のネットワーク構造などによってどう変わる?

2 方法：シミュレーション

下記の条件で架空の母集団からスノーボールサンプリングし、比率の誤差の大きさを推定。Jリーグ引退者に占める自営業者の比率（自営率）を推定するとする。

- 母集団（Jリーグ引退者）規模 1000 人
- 母集団に占める自営業率 pは 0.03, 0.1, 0.3, 0.5
- 母集団の構成員には $i = 1$〜 1000 のIDを振り、1〜$p\times$ 1000を自営、それ以外はその他とする。
- 母集団はスモールワールド (Watts and Strogatz 1998) と仮定し（無向グラフ、平均知人数 =6）、edge の架け替え率 qは 0.02, 0.1, 0.25, 0.5
- $q = 0$ （レギュラーグラフ）時には、i は $i-3, i-2, i-1, i+1, i+2, i+3$ と知人（id の両端数人は例外、図 2.1 参照）

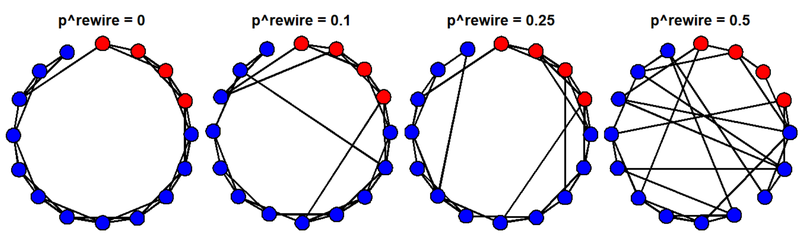

Figure 2.1: スモールワールド・ネットワークの例

- サンプルサイズは 5, 10, 20, 50, 100 人
- 最初の対象者は $i = 1$の人（自営でその他から最遠）、または母集団で最も知人の多い人（複数いたらID番号の一番小さい人）
- 対象者は知人の中から無作為に一人選んで次の対象者として紹介
- 推定方法：
  1. 標本比率 $\hat{p}$から母比率（以下、真値）p を推定し、二項分布でその信頼区間を推定
  2. 切片のみの空間誤差相関モデルで、知人同士に誤差相関があると仮定 (Ward and Gleditsch 2008)。R の spatialreg パッケージで計算 (Bivand, Pebesma, and Gomez-rubio 2013)
- すべての条件組み合わせで 20 回抽出を繰り返し、誤差の大きさ$\sqrt{\frac{\sum(p-\hat{p})^2}{\text{繰り返し数}}}$と95%信頼区間に真値が入る率を計算。どのような条件下で誤差が小さくなるか検討

3 結果

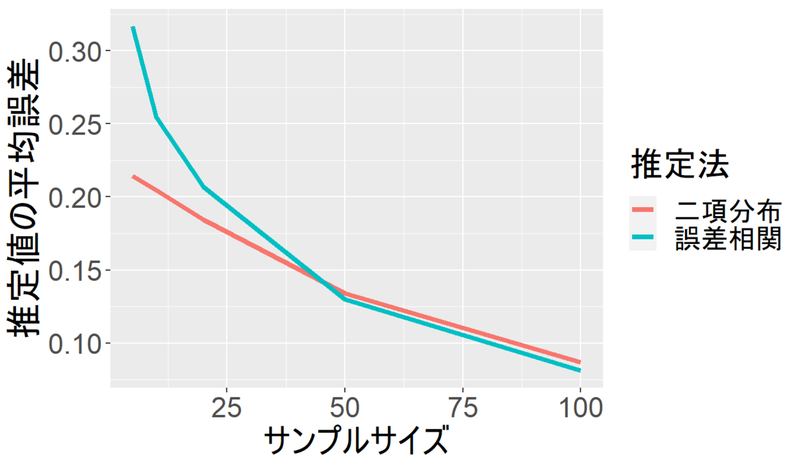

Figure 3.1: 推定法別、サンプルサイズと誤差

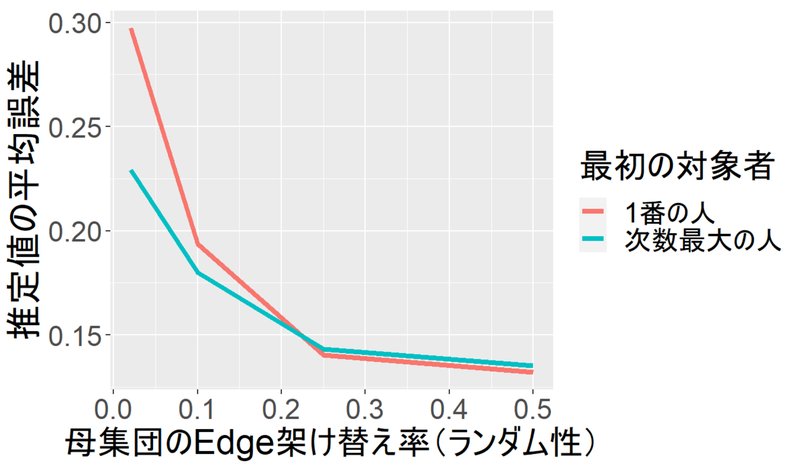

Figure 3.2: 最初対象者・N別平均誤差

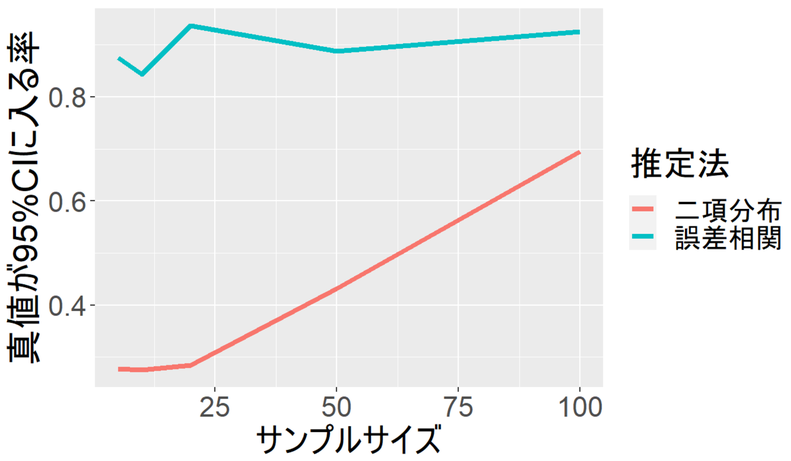

Figure 3.3: 推定法別サンプルサイズと真値が95%信頼区間に入る率

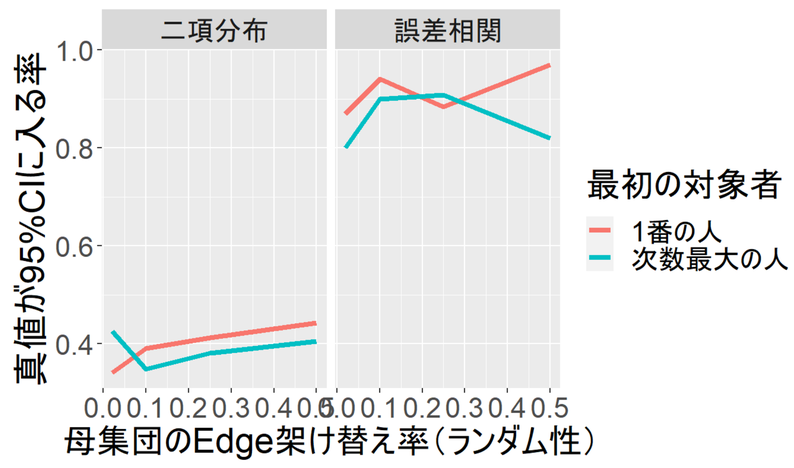

Figure 3.4: 推定法、最初の対象者、架け替え率別真値が95%信頼区間に入る率

*OLS*で誤差予測

	Model 1
切片	0.42
	(0.22)
次数最大者からスタート	-0.19
	(0.12)
誤差相関推定	0.33**
	(0.12)
Nの平方根	-0.23***
	(0.02)
pの平方根	3.87***
	(0.28)
架け替え率の対数	-0.42***
	(0.05)
R^2	0.56
Num. obs.	320

*** p < 0.001; ** p < 0.01; * p < 0.05

4 議論・課題

- どの知人を紹介するかはランダム（等確率）であることが結果に影響？ しかし、架け替え率に還元可能?
- 違うネットワーク構造（例えばスケールフリー）でどうなるかは今後の課題
- 違う中心性の高い人からスタートすれば誤差を減らせる?
- 空間誤差相関モデルは残差が正規分布から離れるほど誤差が増える傾向。誤差相関 probit も課題
- 割愛した交互作用も（次数最大の人からスタートしても、自営率が高い場合や架け替え率が高い場合はむしろ誤差増大）

文献

Bivand, Roger S., Edzer J. Pebesma, and Virgilio Gomez-rubio. 2013. *Applied Spatial Data Analysis with r*. 2nd ed. Springer.

Ward, Michael D., and Professor Kristian Skrede Gleditsch. 2008. *Spatial Regression Models*. Thousand Oaks: Sage.

Watts, Duncan J., and Steven H. Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature* 393(6684):440–42. doi: 10.1038/30918.

太郎丸博. 2021. "有意抽出時の標本誤差とサンプルサイズ、母分散の関係." in 第71回数理社会学会大会報告要旨集.

posterdown