arXiv:2302.00951v1 [stat.AP] 2 Feb 2023

存在报告问题的当前时长数据的贝叶斯分析：利用中低收入国家横断面调查中收集的距上次性行为时间数据估计性行为间隔时间分布

Chi Hyun Lee、Herbert Susmann 和 Leontine Alkema *†

美国马萨诸塞州阿默斯特，马萨诸塞大学阿默斯特分校公共卫生与健康科学学院生物统计学与流行病学系。

摘要

全球各地的人群均使用计划生育综合指标来监测避孕方法的需求和使用情况，例如作为 FP2030倡议的一部分。中低收入国家的计划生育指标通常基于横断面住户调查收集的数据。近期提出的指标通过评估目标人群的性行为间隔时间（TBS）分布，将性活动纳入考量。

本文提出一种统计方法，利用中低收入国家通常可获得的数据估计性行为间隔时间分布，同时应对两项主要挑战。第一项挑战是，有关性行为时间的信息通常仅限于横断面调查中收集的女性距上次性行为时间（TSLS）数据。在我们提出的方法中，我们采用当前时长法，利用现有的TSLS数据估计性行为间隔时间分布，并据此推导群体层面的性行为频率。此外，观测到的TSLS数据存在报告问题，因为报告单位可能不同，也可能经过取整。为应用当前持续时间法并处理这些数据报告问题，我们构建了一个灵活的贝叶斯模型，并对所提出的建模方法作出详细的技术说明。

关键词：贝叶斯推断；当前持续时间法；人口与健康调查；计划生育；报告问题；性活动

*本研究获比尔及梅琳达·盖茨基金会资助（INV-008441）。根据该基金会的资助条件，因本次投稿可能产生的作者接收稿版本已获知识共享4.0通用许可协议授权。

†通讯地址：Chi Hyun Lee，马萨诸塞大学阿默斯特分校公共卫生与健康科学学院生物统计学与流行病学系，阿默斯特，马萨诸塞州01003，美国。电子邮箱：chihyunlee@umass.edu

1 引言

计划生育汇总指标用于监测全球各人群对避孕方法的需求及其使用情况，例如作为FP2030倡议[1]的一部分。中低收入国家的计划生育指标通常基于通过横断面住户调查收集的数据，例如人口与健康调查项目 [1]。最近提出的指标通过评估目标人群中性行为间隔时间（TBS）的分布来衡量性活动 [2]。本文介绍我们用于估计性行为间隔时间分布的统计方法，以全面量化该人群中女性的性活动。

在我们所关注的应用中，估计 TBS 分布面临一些挑战。首先，我们无法直接获得目标人群的 TBS 数据。相反，横断面调查会收集距上次性行为时间（TSLS）数据，并询问参与者“您上一次性交是什么时候？”，这只能提供有关女性性活动的有限信息。TSLS，即当前持续时间，指从最近一次性交到调查时点之间的时间。在我们提出的方法中，我们应用当前持续时间法，利用观测到的 TSLS 数据 [3]推断未观测到的 TBS。此外，TSLS 数据还存在多种报告问题。在横断面调查中，女性可能会以不同单位报告 TSLS，例如天、周、月和年。此外，受访者可能倾向于报告某些特定数值，例如 7 或 30 的倍数。因此，数据由精确报告值和一些粗略值混合构成，导致数据出现“堆积”现象。我们采用贝叶斯建模方法来处理这些报告问题。

本文详细介绍了如何使用灵活的贝叶斯模型，根据 TSLS 数据估计性行为间隔时间密度的建模方法。我们在第 2 节介绍符号和模型，随后在第 3 节简要说明基于当前持续时间法的 TBS 分布与 TSLS 分布之间的关系。在第 4 节，我们介绍所提出的贝叶斯估计方法，用以处理横断面调查数据中的报告问题。最后，我们提供有关计算的信息。

2 符号与方法概述

令 X 表示从目标人群中随机抽取的一名女性的离散性行为间隔时间（以天计）。当离散时间为 d 天时，即 $X = d$，我们假设潜在的（连续）性行为间隔时间处于区间（d，$d+1$]。例如，$X = 0$ 指不超过 24 小时的性行为间隔时间。我们的目标是估计 TBS 的分布，即 $f_X(x) = \text{Prob}(X = x)$，其中 $X \sim f_X(x)$，它表示群体层面的性行为频率。

在我们关注的应用中，TBS 数据无法直接观测。因此，我们使用从人群中的女性收集到的 TSLS 数据。假设我们从人群中随机抽取了 n 名女性。对于第 i 位女性，令 y_i 表示 $i = 1$、2、.....、n 中 TSLS 的精确天数。我们假设 $y_i \sim f_Y(y)$，其中 $f_Y(y) = \text{Prob}(Y = y)$ 表示 TSLS 的分布。TBS 的分布 $f_X(x)$ 可根据有关性交频率的假设，利用估计得到的 TSLS 分布进行估计。$f_X(x)$ 与 $f_Y(y)$ 之间关系的详细说明见第 3 节。第 4 节介绍了用于在此情境下估计 TSLS 分布的灵活的贝叶斯模型。

TSLS 数据可能以不同单位报告。对于第 i 位女性，观测数据由 (z_i、$u[i]$) 构成，其中 z_i 表示报告的 TSLS 值，$u[i]$ 表示报告单位，$u[i] = 1$、2、3、4 分别表示天、周、月和年。我们使用观测数据 $\boldsymbol{y} = \{(z_i$、$u[i])\ i = 1$、.....、$n\}$，在考虑报告误差的基础上估计 TSLS 分布 $f_Y(y)$，具体方法见第4.1节。

3 TBS 与 TSLS 分布之间的关系

假设从目标人群中随机抽取女性，并对每位女性，从其各次性行为事件之间的间隔时间中随机抽取一个间隔时间。此外，假设人群中女性的性行为事件不存在系统性模式

[1]参见 https://fp2030.org/

。U 在这些假设下， $f_X(x)$ 与 $f_Y(y)$ 之间可得出如下关系 [4, 3] :

$$f_Y(y) = \frac{S_X(y)}{E_X(X)}, \tag{1}$$

其中， $S_X(x) = \text{Prob}(X \geq x)$ 表示与 X 相关的生存函数，而 $E_X(X)$ 是 X 的期望值。根据公式 (1)， $f_Y(y)$ 必须是递减分布。给定 $f_Y(y)$，可推导出离散分布 $f_X(x)$，如下所示：

$$f_X(x) = S_X(x) - S_X(x+1) = \frac{f_Y(x)}{f_Y(0)} - \frac{f_Y(x+1)}{f_Y(0)},$$

其中， $f_Y(0) = 1/E_X(X)$。

在实际应用中，通常会指定一个固定常数值 $y_0 < \tau$，其中 $\tau < \infty$ 是观测到的最大 y，以便将分析限制在满足 $0 \leq y < y_0$ 的 y 范围内，并排除任何长于 y_0 的持续时间。这样可以确保数据质量和/或细化目标人群。在研究性活动时，我们排除了过去 2 年内没有过性交的人。我们将 y 的支撑集限制为 $0 \leq y < y_0$，方法是在估计 $0 \leq y < y_0$ 的 $f_Y(y)$ 时施加额外约束 $f_Y(y_0) = 0$，或等价地施加 $S_X(y_0) = 0$。这保证了可推导出在其支撑集 $0 \leq x < y_0$ 上不受限制的 TBS 分布 $f_X(x)$。对于我们的应用，将 $S_X(y_0) = 0$ 设为适当值，因为 TSLS 的经验生存曲线在 2 年时接近 0。

4 TSLS 分布的灵活贝叶斯设定

我们使用灵活的贝叶斯模型来估计非递增离散分布 $f_Y(y)$。为简化符号，令 $\phi_d = f_Y(d) = \text{Prob}(Y = d)$ 表示 TSLS 等于 d 天的概率。根据定义， ϕ_d 必须介于 0 和 1 之间，且总和为 1：$\sum_{d=0}^{D} \phi_d = 1$ 其中 $D = 729$ 天（即 2 年）。为纳入这些约束，我们在变换后的尺度上定义 ϕ_d。具体而言，我们将向量 $\boldsymbol{\phi} = (\phi_0, \phi_1, \ldots, \phi_D)$ 定义为向量 $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_D)$ 的变换，如下所示：

$$\boldsymbol{\phi} = \boldsymbol{\gamma} / \sum_d \gamma_d.$$

经过此变换后，只要满足 $\boldsymbol{\gamma} \geq 0$，向量 $\boldsymbol{\phi}$ 就是一个定义良好的概率集合（即所有概率均介于 0 和 1 之间，且向量元素之和为 1）。

我们使用积分 B 样条 [5] 对 $\boldsymbol{\gamma}$ 建模。 γ_d 的模型设定如下: :

$$\gamma_d = \sum_{k=1}^{K} \alpha_k BI_k(d),$$

其中， $BI_k(d)$ 表示在第 d 天取值的第 k 个积分 B 样条， $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots\ldots, \alpha_K)$ 为样条系数。所用的积分 B 样条见图1（A）。我们对 α_k 进行如下建模:

$$\begin{aligned} \alpha_k &= \exp\left(\sum_{j=k}^{K} \delta_j\right), \\ \delta_j | \sigma &\sim N(0, \sigma^2), \\ \sigma &\sim N^+(0, 1). \end{aligned}$$

在此设定下，所得的 γ_d 为正值，且随 d 增大而递减。图1 (B) 和 (C) 分别展示了 $\delta_j = 0$ 对所有 j、 $\alpha_k = 1$ 对所有 k 时的 TSLS 和 TBS 分布。

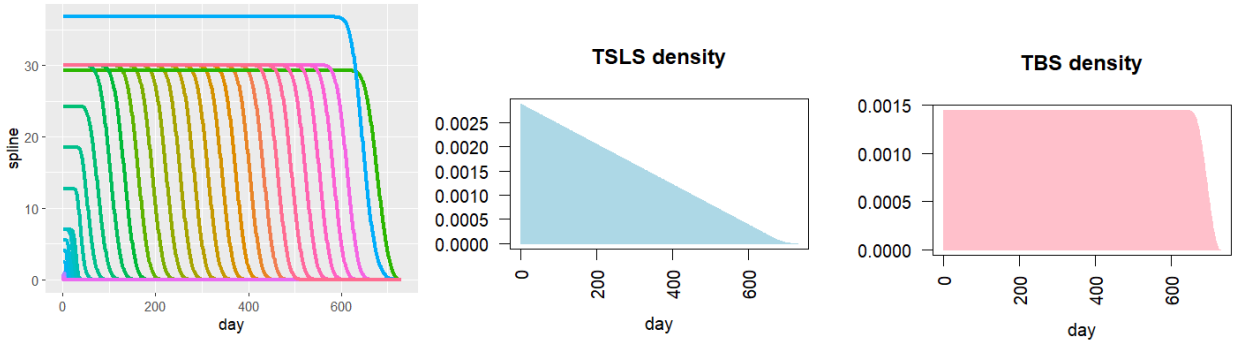

图1：**样条函数以及当样条系数设为** $\alpha_k = 1$ **对所有** k**时得到的 TSLS 和 TBS 分布示意图**。(A) 用于估计转换后的 TSLS 概率的积分 B 样条。每条彩色线表示一个样条函数 $BI(\cdot)$，横轴为天数。(B) 对应的 TSLS 分布，其中 $\alpha_k = 1$。(C) 对应的 TBS 分布。

4.1 报告假设

我们观测到报告的 TSLS 值 z_i，以及报告单位 $u[i] = 1$，2，3，4，分别对应天、周、月或年。对于以周为单位报告的数据，我们对 TSLS 的精确值 y_i 作如下假设：

$$y_i \in \{z_i \cdot 7, \ldots, (z_i + 1) \cdot 7 - 1\}, \text{ for } u[i] = 2,$$

对于以月为单位报告的数据，我们假设：

$$y_i \in \{z_i \cdot 30 + 1, \ldots, (z_i + 1) \cdot 30\}, \text{ for } u[i] = 3.$$

对于以年为单位报告的数据，如果 TSLS 的报告值为 $z_i = 1$，具体而言，我们假设：

$$y_i \in \{365 - 31, \ldots, 729\}, \text{ for } u[i] = 4.$$

对于以天为单位报告的数据，我们假设受访者可能倾向于按 7 天或 30 天的倍数报告，也可能准确报告非 7 天或 30 天倍数的天数。我们按如下方式纳入这一假设：

$$\begin{aligned} &y_i = z_i \text{ for } u[i] = 1 \text{ and } z_i \notin \{7, 14, 21, 28, 30, 60, 90\}, \\ &y_i \in \{z_i - 2, \ldots, z_i + 2\} \text{ for } u[i] = 1 \text{ and } z_i \in \{7, 14, 21, 28, 30, 60, 90\}. \end{aligned}$$

基于这些报告假设，似然函数总结如下：

$$\begin{aligned} z_i &\sim f_Z^{(day)}(\cdot) \text{ for } u[i] = 1 \text{ (reporting in days)}, \\ z_i &\sim f_Z^{(week)}(\cdot) \text{ for } u[i] = 2 \text{ (reporting in weeks)}, \\ z_i &\sim f_Z^{(month)}(\cdot) \text{ for } u[i] = 3 \text{ (reporting in months)}, \\ z_i &\sim f_Z^{(year)}(\cdot) \text{ for } u[i] = 4, z_i = 1 \text{ (reporting in years)}, \end{aligned}$$

其中

$$f_Z^{(day)}(z) = \begin{cases} f_Y(z) & \text{for } z \notin \{7, 14, 21, 28, 30, 60, 90\}, \\ \sum_{y=z-2}^{z+2} f_Y(y) & \text{for } z \in \{7, 14, 21, 28, 30, 60, 90\}, \end{cases}$$

$$f_Z^{(week)}(z) = \sum_{y=z\cdot 7}^{(z+1)\cdot 7-1} f_Y(y),$$

$$f_Z^{(month)}(z) = \sum_{y=z\cdot 30+1}^{(z+1)\cdot 30} f_Y(y),$$

$$f_Z^{(year)}(1) = \sum_{y=365-31}^{729} f_Y(y).$$

将报告数据绘制为直方图时，我们为与 z_i 关联的 y_i 集合赋予相等的概率质量。

4.2 计算

我们开发了一个 R 软件包，可在 GitHub 上通过 https://github.com/AlkemaLab/safpet 获取。拟合模型时，我们使用 R 和 Stan [6, 7, 8]，采用 HMC 算法从定义 TSLS 分布的模型参数的后验分布中进行抽样。利用贝叶斯估计模型，我们获得了所有模型参数的一组后验样本。根据样条系数 $\alpha_k^{(s)}$ 在 $s = 1, \ldots, S$ 的后验样本，我们得到估计的 TSLS 和 TBS 分布；其中 s 表示后验样本索引。具体而言，后验样本的每一种组合（$\alpha_1^{(s)}, \ldots, \alpha_K^{(s)}$）都会生成一个 TSLS 和 TBS 联合分布。

R 和 Stan 代码用于生成我们近期应用研究的结果 [2]。在该研究中，我们运行了四条并行链，每条链共迭代 2,000 次。每条链在预热阶段的前 1,000 次迭代均予以舍弃，因此最终每条链包含 1,000 个样本。我们使用标准诊断检查来检查收敛情况和采样效率。这些检查基于轨迹图、使用秩正态化抽样值的改进版 Rhat 诊断（[9], [10]），以及各种有效样本量（ESS）计算，包括主体有效样本量和尾部有效样本量——后者取 5% 和 95% 分位数有效样本量中的较小值。

参考文献

[1] Trevor N. Croft、Aileen M. J. Marshall、Courtney K. Allen 等。《DHS 统计指南》。ICF，美国马里兰州罗克维尔，2018 年。[2] Chi Hyun Lee、Herbert Susmann 和 Leontine Alkema。基于性活动和避孕措施使用数据的计划生育新指标及妊娠风险暴露。《计划生育研究》，2023 年。即将刊出。[3] Niels Keiding、Kajsa Kvist、Helle Hartvig、Mads Tvede 和 Svend Juul。根据横断面样本中的当前持续时间估计受孕时间。《生物统计学》，3(4):565–578，2002 年。[4] Paul D. Allison。回溯复发时间的生存分析。《美国统计学会杂志》，80(390):315–322，1985 年。[5] J. O. Ramsay。单调回归样条的应用。《统计科学》，3(4):425 – 441，1988 年。[6] Bob Carpenter、Andrew Gelman、Matthew D. Hoffman、Daniel Lee、Ben Goodrich、Michael Betancourt、Marcus Brubaker、Jiqiang Guo、Peter Li 和 Allen Riddell。《Stan：一种概率编程语言》。《统计软件杂志》，76(1)，2017 年。[7] R 核心团队。《R：一种用于统计计算的语言与环境》。R 统计计算基金会，奥地利维也纳，2021 年。[8] Stan 开发团队。Stan 建模语言用户指南和参考手册，2.18.0 版，2018 年。[9] Andrew Gelman 和 B. Rubin。使用多条序列进行迭代模拟推断。《统计科学》，7(4):457–72，1992 年。[10] Aki Vehtari、Andrew Gelman、Daniel Simpson、Bob Carpenter 和 Paul-Christian B¨urkner。秩正态化、折叠与局部化：一种用于评估 MCMC 收敛情况的改进 R hat。贝叶斯分析，2020 年 7 月。